

LA ESTRUCTURA DE LOS DATOS

Otra peculiaridad de los datos científicos es que, pese a la etimología de la palabra “dato”, no son nada dado, sino que hay que producirlos, y muchas veces, laboriosamente (Bunge; 1979)

“¿Qué es un dato?” es la pregunta a la que quisiéramos responder en este primer capítulo. Intentaremos mostrar así como la estructura de cualquier dato es esencialmente el producto de un acto clasificatorio, acto cuya simplicidad es sólo aparente. Lejos de sernos inmediatamente “dado” –como lo sugiere engañosamente la etimología-, el dato presupone toda una serie de operaciones que se llevan a cabo en forma simultánea dentro de un sistema conformado por múltiples dimensiones. De este modo, el dato más elemental ya conlleva toda una serie de elecciones teórico-metodológicas. Y es que desde el punto de vista que se sostiene aquí, el dato es eminentemente racional: no se “ve” más que lo que ya se está dispuesto a ver.¹ Comenzaremos procediendo a un ejercicio de disección de la naturaleza del dato, mediante el cual estaremos en condiciones de aislar sus elementos constitutivos. La exposición algo dogmática de una serie de definiciones nos permitirá establecer las relaciones sistemáticas por las que se vinculan unos a otros los conceptos de “sistema” y “unidad de análisis”, “propiedades” y “medición”, “variables” y “valores”. Para ello, nos valdremos de ejemplos muy sencillos, deliberadamente no sociológicos. Al término de este trabajo analítico, el dato deberá aparecer como el producto de la imbricación de dos estructuras: la que permite atribuir una propiedad a un objeto, y la que posibilita estructurar sistemáticamente conjuntos de propiedades.

1. UNIDADES DE ANÁLISIS, VARIABLES, VALORES

Según nos dice Galtung², todo dato hace referencia a una estructura constituida por tres elementos: **unidad de análisis, variable y valor**. Cualquier dato aislado consistirá pues en...

a) una **unidad de análisis** que

b) en una **variable** específica presentará

c) un determinado **valor**.

Lo que constituye el dato son estos tres elementos considerados conjuntamente con las relaciones que mantienen entre sí. Sea cual sea nuestro objeto de estudio, si queremos sostener cualquier proposición empírica acerca de él, se lo deberá concebir en términos de esa estructura tripartita. “Unidad”, “variable” y “valor” son todos términos que denotan determinados tipos de conceptos. Una sencilla aproximación desde la lógica puede permitirnos echar algo más de luz sobre su significado, partiendo de las categorías clásicas de “sujeto” y “predicado”. La idea más elemental de lo que es un dato puede ejemplificarse a partir de un enunciado muy simple: “ese sillón es rojo”. Se puede reconocer en este enunciado un sujeto (“sillón”) y un predicado (“rojo”); también podemos decir que al hacer esta afirmación estamos denotando una determinada **propiedad** de un objeto. Lo mismo sucede cuando sostengo, por ejemplo, que “Rómulo es argentino”; “argentino” es una propiedad que se predica del sujeto de la oración (“Rómulo”). Ahora bien, las fórmulas “ese sillón es rojo” y “Rómulo es argentino” consisten simplemente en enunciados que se refieren respectivamente a diferentes proposiciones³. Por lo tanto, es posible establecer desde ya que:

Un dato se expresa en una proposición⁴

Togerson (1965) propuso el uso del término “sistema” para denotar todo tipo de objetos o de “cosas”; así, en nuestros ejemplos, tanto “ese sillón” como “Rómulo” serían sistemas. Es evidente que, por definición, la variedad de sistemas que pueden distinguirse no tiene límites: una célula es un “sistema” y también lo son un auto, un alumno, un libro o una galaxia. A su vez, cada sistema puede ser caracterizado por toda una serie de propiedades:

un auto en particular podrá ser *rojo*,
caro,
veloz,
antiguo,
etc.

Sin embargo, observemos desde ya que no todas las propiedades son aplicables a todos los sistemas. No tendría sentido pensar en un libro veloz, así como carecería de significado hablar de una galaxia “cara” (al menos para nosotros aquí y ahora). Esta es una simple consecuencia del hecho de que un **dato** debe expresarse en una **proposición**, vale decir en algo que necesariamente habrá de ser **verdadero o falso**. Así el enunciado “este libro es veloz” no haría referencia a ninguna proposición.⁵

Cuando a partir de la observación de un sistema predicamos una propiedad podemos decir que estamos realizando una medición.⁶ Lo fundamental es que los sistemas no son jamás medidos en sí mismos sino que lo que medimos son siempre propiedades de los sistemas; es así que estamos en condiciones de producir una nueva definición:

Un dato es el producto de un procedimiento
De medición, y medir supone predicar una propiedad.

Encontramos ejemplos de datos muy simples en enunciados del tipo

“Este auto es*rojo*”.
“Esta galaxia es*lejana*”.
“Este libro pesa.....*358 gramos*”.

Es desde este punto de vista que se puede afirmar que el conocimiento científico consiste en la identificación de sistemas de determinadas clases, en la medición de sus propiedades, y en el establecimiento de relaciones entre dichas propiedades. Por lo demás, es obvio que en ciencias sociales no nos interesan todos los sistemas, ni tampoco todas sus propiedades. De hecho, sucede que precisamente puede entenderse que lo que diferencia a las disciplinas científicas unas de otras son en gran medida los tipos de sistemas de los que se ocupan y las propiedades de dichos sistemas que toman en cuenta.⁷

Por otra parte, basándonos en determinadas propiedades podemos referirnos a clases de sistemas:

las **galaxias lejanas**,
los **autos 0 kilómetro**,
los **libros de Sociología**,
los **alumnos de Metodología**,
etc.

También de estas clases de sistemas podemos predicar propiedades:

“**Los autos 0 km.** son *caros*”,
“**Las galaxias lejanas** son *difíciles de observar*”,
“**Los libros de Sociología** son *aburridos*”,
“**Los alumnos de Metodología** son *sabios*”,
etc.

Así como existen autos de diversos colores también existen lápices, flores, libros y cielos de distintos colores. Todos concordaremos sin embargo en que un auto no es ni un lápiz ni una flor, ni un libro, ni un cielo. Si queremos saber qué es un auto podemos recurrir al Larousse usual: "Vehículo que camina movido por un motor de explosión". Todos coincidiremos en que un lápiz no "camina", en que una flor no es un "vehículo" y en que los libros y cielos no son "movidos por un motor de explosión". Es imposible que confundamos un auto con un lápiz, porque de algún modo poseemos **definiciones** de estos objetos; sabemos que un auto se define por determinadas características: tener cuatro ruedas, un volante, asientos y una carrocería, y ser capaz de llevarnos con relativa –y cada vez menor- seguridad de un punto a otro de la superficie terrestre, rodando por los caminos. En otras palabras, el sistema "auto" se caracteriza por presentar determinadas propiedades, a saber:

ser *vehículo*,

caminar,

ser movido por un *motor a explosión*,

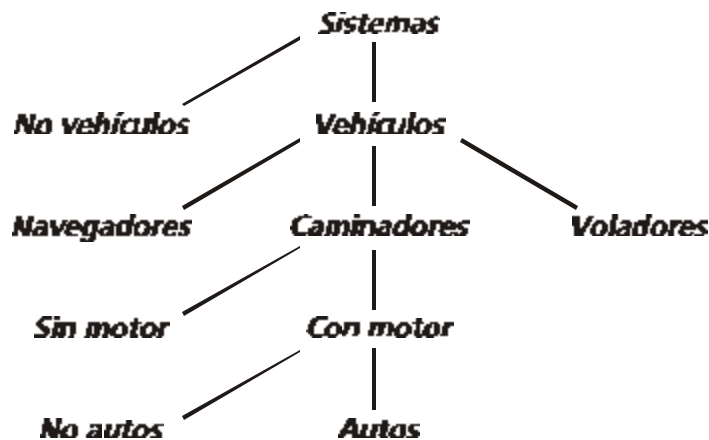
tener *cuatro ruedas*,

etc.

Se pueden imaginar fácilmente sistemas que no satisfagan algunas de estas propiedades: una avioneta Piper o una lancha no "caminan", una bicicleta no es "movida por un motor de explosión", etc...

Lo que queremos remarcar, entonces, es que no solamente existen sistemas de los cuales es posible predicar propiedades, sino que los sistemas mismos son susceptibles de ser definidos en última instancia a partir de una determinada combinación de propiedades. Mucho "antes" de ser *rojo*, un auto es **auto** en tanto presenta determinadas propiedades. Podríamos pensar que todos los sistemas se dividen en dos **clases**: "Vehículos" y "No-vehículos"; que algunos de los vehículos "caminan" mientras otros no lo hacen, y que sólo algunos de estos últimos son "movidos por un motor de explosión":

Figura I: Tipos de sistemas



A partir de las nociones de “sistema” y de “propiedad”, podemos establecer los conceptos de “universo” (y, correlativamente, el de “muestra”), ⁸“unidad de análisis”, “variable” y “constante”. Estos términos cobrarán un sentido específico cuando se los considere dentro de una estructura definida en un determinado nivel de inclusión. Así, podemos reconocer en el “**Universo**” el nivel de mayor inclusión posible, que se compone de todos los sistemas existentes, de todos los que han existido y existirán; en este nivel se incluye también –al menos potencialmente- todas las propiedades posibles. Ahora bien, apelando a determinadas propiedades podemos definir sub-universos –o, simplemente “**universos**”- dentro de este “Universo” mayor en un número ilimitado. Por ejemplo, es posible definir el universo de los “vehículos. Los vehículos son sistemas que permiten desplazar a otros sistemas (y también a sí mismos) de un punto a otro del espacio; se los define por poseer la propiedad de ser “medios de locomoción”. ⁹ Se los puede distinguir así de todos los otros sistemas en los que no reconocemos la propiedad de funcionar como medios de locomoción. Todos los vehículos comparten esta propiedad de ser “medios de locomoción”; para esta clase de sistemas, el ser “medio de locomoción” es una **constante**. Notemos que en el nivel de inclusión inmediatamente superior –el Universo- no todos los sistemas son “medios de locomoción”: en ese nivel superior dicha propiedad funciona como una **variable**. Podemos continuar definiendo universos de menor nivel de inclusión. Así tendremos el universo de los vehículos “caminadores” (que se diferencia de los universos de los vehículos “voladores” y “navegadores”); a su vez, dentro del universo de los vehículos caminadores, definimos el de los “movidos a motor”, y dentro de éste, el universo de los autos. La Figura II permite apreciar cómo cada universo ¹⁰ está definido por una serie de propiedades constantes y se corresponde con un determinado nivel de inclusión. Todo nivel de inclusión inferior requiere para su definición a partir del nivel inmediatamente superior tomar como constante al menos una propiedad; y viceversa: al pasar de un nivel de inclusión menor a otro superior aumentará el número de propiedades variables.

Figura II: Ejemplo de definición de universos sobre la base de Propiedades constantes y variables

UNIVERSOS	PROPIEDADES			
	Medio de locomoción	Medio en que se desplaza	Medio de impulsión	Número de ruedas
Sistemas	Variable	Variable	Variable	Variable
Vehículos	Constante	Variable	Variable	Variable
Vehículos caminadores	Constante	Constante	Variable	Variable
Vehículos caminadores a motor	Constante	Constante	Constante	Variable
Autos	Constante	Constante	Constante	Constante

En cualquier nivel de inclusión que se tome, las propiedades aparecerán Jugando entonces un doble papel: algunas de ellas, tomadas como constante, estarán definiendo un universo, mientras que otras funcionarán como variables y podrán ser objeto de investigación. Podemos ahora introducir la idea de que todos los sistemas que corresponden a un nivel de inclusión dado conforman un universo, y que cada uno de estos sistemas que componen dicho universo es susceptible de ser considerado como una unidad de análisis. Ello nos permite definir:

Una unidad de análisis es un sistema definido por presentar determinadas propiedades, algunas de ellas constantes (las que definen su pertenencia a un universo compuesto por todos los sistemas que presentan esas mismas propiedades) y otras variables (las que podrán ser materia de investigación dentro de ese universo).

Disponemos ahora de una primera definición de lo que es una “variable”, entendida como un determinado tipo de propiedad, a partir de los conceptos de “unidad de análisis” y de “universo”. En lo que sigue, nos proponemos continuar elucidando este concepto de “variable”, para lo cual deberemos introducir una nueva idea: la de “valor”.

La siguiente definición de Galtung es un buen punto de partida para esclarecer las relaciones que median entre los tres elementos constitutivos del dato:

“Dado un conjunto de unidades, un valor es algo que puede predicarse de una unidad, y una variable es un conjunto de valores que forma una clasificación.”¹¹

De acuerdo a lo que se ha venido exponiendo, no debe haber ninguna dificultad en reconocer de inmediato que un “valor” es simplemente una propiedad. Ahora bien, hemos dicho que una variable también es una propiedad: ¿Significa esto que “valor” y “variable” son dos términos que denotan el mismo concepto, el de “propiedad”? En absoluto.

Cuando enunciamos “Este auto es *rojo*”, ¿En qué sentido es posible hablar de “medición”? Solamente en tanto esta propiedad “*rojo*” se encuentra definida por su lugar dentro de una estructura que le confiera un sentido. Así la estructura más simple en la que podemos pensar se basa en una relación simple de oposición:

rojo versus *no-rojo*

De este modo afirmar que el auto es rojo equivale a afirmar que dicho auto **no** es no-rojo: la afirmación de la propiedad es al mismo tiempo la negación de la no-propiedad. Podríamos introducir un término para denotar esta propiedad —el de “*rojidad*”— que denotaría una variable, entendida ahora como una característica susceptible de adoptar diferentes valores (“sí” y “no”). Cualquier unidad de análisis de la cual se pudiera predicar con sentido la rojidad presentaría en esta variable el valor “sí” o el valor “no”. Así como podríamos medir la longitud de un auto (“este auto mide 427 cm.”), así también mediríamos su “rojidad”:

“Este **auto** presenta el valor “sí” en la variable *rojidad*”
(o, para decirlo más simplemente: “Este auto es rojo”)

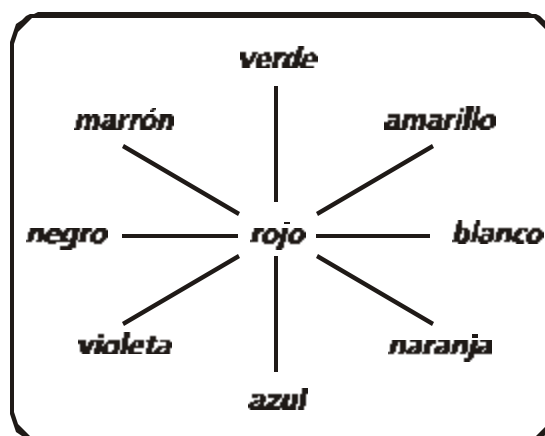
De este modo, podemos establecer que:

una variable es un conjunto estructurado de valores¹²

Ahora bien, el término “rojo” también puede definirse por su pertenencia a una estructura algo más compleja; dentro de esta estructura, el “rojo” se definirá como tal por no ser ni marrón, ni verde, ni amarillo, ni blanco, ni naranja, ni azul, ni violeta, ni negro. “Rojo” puede aparecer ahora como un

valor de una variable a la que podríamos denominar “color” y en la que se opone a toda una serie de otras propiedades que constituyen otros tantos valores posibles de esa variable.

Figura III: Ubicación de “rojo” dentro de una estructura de oposiciones



Los nombres de colores incluidos en la Figura III constituyen “un conjunto de valores que forma una **clasificación**”, vale decir, lo que Galtung define como “**variable**”. “Variable” y “valor” aparecen pues como conceptos bien distintos, aunque cuando se encuentren estrechamente relacionados. Un valor es parte de una variable, no existe sino por las relaciones que guarda con los otros valores que componen esta variable. A su vez, una variable no es otra cosa que el conjunto de los valores que la conforman y de las relaciones que éstos mantienen entre sí. Las variables correctamente construidas deben necesariamente satisfacer los requisitos demandados a cualquier clasificación: los valores que las compongan deberán en cada caso ser **exhaustivos** -en su conjunto-, y **mutuamente excluyentes**. Así como la hemos construido, nuestra variable “color” es imperfecta: un auto rosado no encontraría ubicación en ningún valor; nuestra clasificación no es exhaustiva, y requeriría del agregado de otros valores (o **categorías**). En cambio, sí resultan mutuamente excluyentes las categorías (o valores) que la componen: cualquier persona de cuatro o más años en nuestra cultura dispone de un sistema de reglas que le permite aplicar correctamente estos valores.

(Baranger, Denis (s/f) *La Estructura de los Datos*, Capítulo I de su: **Construcción y Análisis de Datos. Una introducción al uso de técnicas cuantitativas en la Investigación Social**, Editorial Universitaria)

¹ Desde la tradición epistemológica francesa (Cf. Bachelard, 1972 y 1986) esto es una obviedad. En la vereda opuesta, el positivismo lógico (CF. Ayer, 1965) sostenía la posibilidad de un lenguaje observacional puro dentro del cual era factible formular enunciados que permitieran dirimir acerca de la verdad –falsedad de una hipótesis. En el mundo anglosajón, hubo que esperar hasta Kuhn (1969) para que se pusiera radicalmente en cuestión la reconstrucción positivista.

² Cf. Galtung, 1968: I, cap. 1.

³ Es importante tener en cuenta esta distinción que establece la Lógica. Así, “ese sillón es rojo” y “*ce fauteuil est rouge*” son ciertamente enunciados diferentes, pero ambos hacen referencia a la misma proposición. Según Copi, “Se acostumbra usar la palabra “proposición” para designar el significado de una oración declarativa (...) Una oración declarativa forma siempre parte de un lenguaje determinado, el lenguaje en el cual es enunciada, mientras que las proposiciones no son propias de ninguno de los lenguajes en los cuales pueden ser formuladas”(1962:21)

⁴ Según Bunge, “Las ideas que expresan el resultado de una fase de observaciones son un conjunto de datos. Un dato es una proposición singular o existencial como, por ejemplo, “Se inyectó a la rata # 13 1 mg de nicotina el primer día” (1979:742)

⁵ “Para los fines de la lógica, puede definirse una proposición como algo de lo cual es posible afirmar que es verdadero o falso” (Cf. Cohen y Nagel, 1968:I, 41). Convendrá notar que esto abre la posibilidad de que existan **datos falsos**; así contrariamente a lo afirmado con anterioridad, lo cierto es que “(mi amigo) Rómulo es uruguayo.”

⁶ al menos implícitamente, puesto que la simple afirmación de la propiedad es lógicamente equivalente a la negación de la no-propiedad.

⁷ O para decirlo de otra manera, las ciencias se diferencian por sus “objetos”.

⁸ Cf. *Infra*, capítulo III.

⁹ Larousse *dixit*.

¹⁰ Excepto, claro está, el primero de todos: el “Universo de todos los sistemas”.

¹¹ Cf. Galtung, 1968: I, 78.

¹² Para Frege, uno de los padres de la moderna Lógica, “Una variable es un signo que, en lugar de tener un designado fijo, “recorre” un espectro de posibles designados, y la determinación de un designado particular consistirá en la asignación de un valor para esa variable” (Cf. Lungarzo, 1986: I, 36).

LA MATRIZ DE DATOS




Ante cualquier objeto de investigación, ya sea éste de significación teórica o de importancia meramente práctica, las decisiones metodológicas propiamente dichas tienen necesariamente lugar dentro de un cierto marco conceptual. Dentro de ese marco se debe determinar: a) el grado en que dicho objeto es susceptible de ser producido en tanto objeto científico dentro de la estructura de la matriz de datos; y b) todas las operaciones que hagan a la producción del objeto dentro de este esquema, y muy particularmente las que tengan que ver con la definición de las unidades de análisis y de las propiedades que les sean aplicables.

En este capítulo, mostraremos primero el modo en que se articulan los conceptos de “unidad de análisis”, “variable” y “valor” bajo la forma de la matriz de datos. Nos encontramos ahora en condiciones de aplicar estos conceptos para desmenuzar el significado empírico de cualquier proposición.¹ Todo el problema consiste precisamente en expresar en el lenguaje de los datos el contenido de un enunciado, por lo que la tarea es propiamente la de lograr una **traducción** adecuada de ese contenido.

Comenzaremos con varios ejemplos de ejercitación de esta operación de traducción para generar en forma práctica la idea de la matriz de datos. Luego nos limitaremos a glosar las partes más pertinentes del texto de Galtung, e introduciremos algunos elementos acerca de modalidades alternativas de diseño de una matriz. Finalmente, la distinción de Lazarsfeld y Menzel entre propiedades individuales y colectivas nos permitirá terminar de dibujar un cuadro somero de las distintas posibilidades abiertas en el proceso de construcción de los datos.

1. LA FORMA DE LA MATRIZ DE DATOS

Si afirmamos: “En 1980, había en la Provincia de Misiones 50.553 hogares con necesidades básicas insatisfechas”,² podemos distinguir en este enunciado:

1. Una unidad de análisis  la “Provincia de **Misiones**”.
2. Una variable  el “*Número de hogares con necesidades básicas insatisfechas*”.
3. Un valor  “50.553”.

Estrictamente, limitándonos a esta sola unidad de análisis, puede parecer poco pertinente hablar de una “variable”, puesto que ésta no variaría. No obstante, el número de hogares con necesidades básicas insatisfechas pudo haber sido otro. De hecho la “variable” es tal en la medida en que hace posible la comparación entre varias unidades de análisis: así, por ejemplo, para ese mismo año de 1980 la unidad de análisis Capital Federal presentaba en esta variable un valor de 67.692.³

Este enunciado es de los más simples, desde luego, como que se refiere a una sola característica de un objeto único. Ocurre que muchas veces estamos interesados en proposiciones que se refieren a toda una clase de objetos, y en las que se establecen relaciones entre varias de sus propiedades. Por ejemplo: “En las elecciones para Gobernador del 6 de septiembre de 1987 en Posadas, las mujeres votaron por los candidatos radicales en mayor proporción que los varones”. Deberíamos distinguir aquí:

1. Varias unidades de análisis  los “**electores**”⁴ del 6 de

septiembre de 1987 en Posadas;

2. Dos variables

➡ 2.1. Sexo;

➡ 2.2. Dirección del voto;

3. Los valores que conforman respectivamente estas variables

➡ 3.1. "1"(Varón)/"2"(Mujer);⁵

➡ 3.2. "1"(Radical)/"2"(No radical).⁶

La matriz es una forma de hacer inmediatamente visible la estructura tripartita de estos datos. Así, suponiendo que se haya trabajado con una muestra de 10 electores, tendríamos:

Figura V: Ejemplo de matriz de 10x2

Unidades de análisis	Variables	
	1. Sexo	2. Voto
01	1	2
02	2	1
03	1	2
04	1	2
05	2	1
06	2	2
07	2	1
08	1	1
09	1	2
10	2	2

Cada fila de la matriz corresponde a una unidad de análisis (identificada por un código de 01 a 10), cada una de las dos columnas a una variable, y en las celdas figuran los valores.⁷ Entre los valores, obtendríamos una proporción de 1/5 votos por el Radicalismo; entre las mujeres habría 4/5 votos radicales. Como $4/5 > 1/5$ deberíamos aceptar como verdadera nuestra hipótesis.⁸

Esta proposición, aun cuando fuera verdadera, sería todavía muy puntual. Si aspiráramos a lograr un conocimiento más general del comportamiento electoral de la población argentina, podríamos tener interés en formular una hipótesis del tipo "En 1987 las mujeres votaron por el Radicalismo en mayor medida que los varones". En este caso las variables y los valores serían los mismos, pero sería conveniente que nuestras unidades de análisis se multiplicaran hasta abarcar electores de todos los distritos del país. En este caso, lo que era una propiedad constante de nuestras unidades de análisis "estar inscripto en el padrón electoral de Misiones" pasaría a funcionar como una variable adicional que podríamos denominar "*Jurisdicción*".⁹ Con una muestra de 500 electores tendríamos:

Figura VI: Ejemplo de matriz de 500 x 3

Unidades de análisis	Variables		
	1. Sexo	2. Voto	3. Jurisdic.
001	1	2	17
002	2	1	21
003	1	2	04
004	1	2	07
005	2	1	08

006	2	2	14
007	2	1	13
008	1	1	01
009	1	2	01
010	2	2	02
499	1	2	24
500	2	1	19

Así, por ejemplo, la unidad de análisis “006” sería una “mujer” que votó “No radical” en “Misiones”. De un modo más general, se puede describir a cualquier matriz de datos como respondiendo a la siguiente estructura:

Figura VII: Forma general de la matriz de datos

	V1	V2	V3	...	Vj	...	Vm
UA1	R11	R12	R13	...	R1j	...	R1m
UA2	R21	R22	R23	...	R2j	...	R2m
UA3	R31	R32	R33	...	R3j	...	R3m
...
Uai	Ri1	Ri2	Ri3	...	Rij	...	Rim
...
UAn	Rn1	Rn2	Rn3	...	Rnj	...	Rnm

Fuente: adaptado de Galtung (1966:1,3)

En cualquier investigación se considerará un número finito “n” de unidades de análisis (UUA) y un número finito “m” de variables (V). Uan será entonces la última (o la “enésima”) unidad de análisis incluida en una investigación.¹⁰ Por su parte, cada una de las variables se compondrá de un número “r” de valores. En el esquema presentado Rij es el valor que presenta la Uai en la variable j. Esta es más precisamente la forma en que aparecen los datos en un archivo de computadora; es por tanto la forma en que deberán estar dispuestos nuestros datos para estar en condiciones de procesarlos electrónicamente.¹¹

Nos encontramos ahora en condiciones de recapitular los tres principios fundamentales¹² que debe satisfacer la construcción de una matriz de datos.

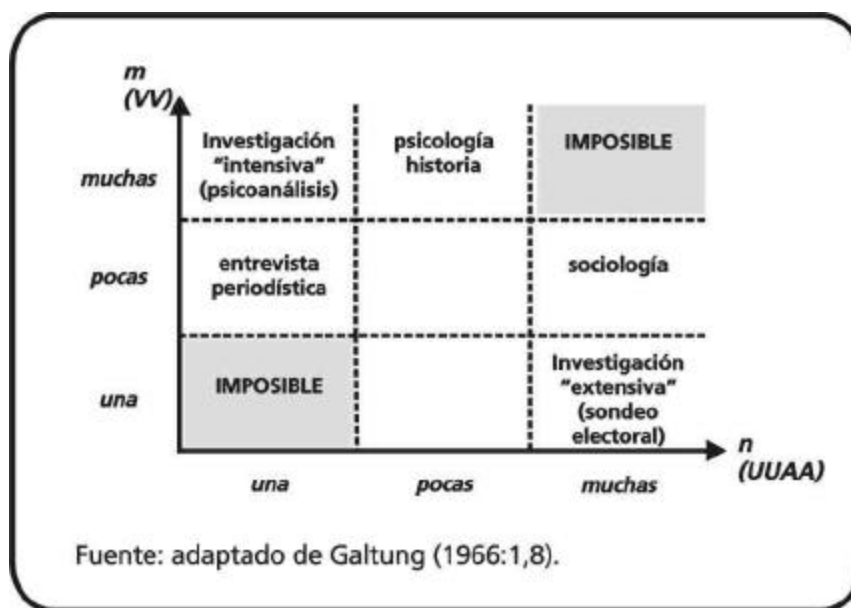
1. Principio de **comparabilidad**: la proposición “Uai Vj da el valor correspondiente en Rk” debe ser verdadera o falsa para cada i, j y k. En otras palabras, a la idea básica de que todas las unidades han de ser medidas en las mismas variables, se agrega la condición previa de que cada una de las combinaciones de una variable determinada con una unidad de análisis debe **tener sentido**: debe ser verdadero o falso que una combinación (UaiVj) presenta un valor determinado Rjk. Por medio de este principio se hacen comparables las variables, las unidades y los valores. La comparabilidad se obtiene cuando las tres series se ajustan las unas a las otras de tal manera que se satisfaga la condición enunciada en este primer principio. Así, si Uai es una nación, Vj la variable “estado civil” y Rjk la lista de los elementos “casados, solteros, viudos y divorciados”, cualquier combinación no será ni verdadera ni falsa, sino que simplemente no tendrá sentido.
2. Principio de **clasificación**: para cada variable Vj la serie de las categorías de respuestas Rjk debe producir una clasificación de todos los pares (UaiVj) (R=1,...n). Para cada variable la serie de sus valores debe formar una clasificación; para cada par UaiVj deberá haber un Rjk (exhaustividad), y sólo uno (exclusión mutua) en que el par pueda ser ubicado. Aplicado a un cuestionario de encuesta, este principio significa que cada interpelado deberá marcar una, y sólo una, respuesta o categoría.¹³
3. Principio de integridad: para cada par (UaiVj) debe hallarse empíricamente un valor Rk. En lo que hace al trabajo empírico de llenado de la matriz, el *desideratum* es no dejar ninguna celda

vacía. En la práctica, se debe intentar que la cantidad de celdas sin información se mantenga lo más baja posible.¹⁴

Desde este punto de vista de la matriz de datos, es factible considerar que las investigaciones pueden diferenciarse según el número de unidades de análisis y de variables que toman en cuenta. Se puede pensar así en investigaciones más o menos “intensivas” –según la cantidad de variables que consideren- y más o menos “extensivas” –de acuerdo con el número de unidades de análisis que sean observadas. Combinando ambas características, generamos una tipología de las investigaciones posibles.

Según Galtung, “Lo ideal es la combinación (muchas, muchas) –tantas unidades y tantas dimensiones como sea posible. Sin embargo, nosotros suponemos que la palabra “muchas” se utiliza de tal manera que esto es imposible, por falta de recursos tales como tiempo, energía, personal y dinero”.¹⁵ Del mismo modo la investigación del tipo (una, una) no tendría sentido. El punteado de las líneas que delimitan las celdas intenta representar la dificultad en establecer un límite rígido entre las distintas cantidades de casos y de variables. El diagrama nos invita a pensar cómo, más allá de su ángulo inferior izquierdo, existe un vasto espacio de posibilidades para realizar investigaciones de naturaleza sumamente diversa y que en cuanto tales demandarán distintas habilidades. Las diferentes celdas representan en definitiva distintas **estrategias** de investigación.

Figura VIII: Tipos de investigación, según el número de variables y de unidades de análisis involucradas



Si la matriz se reduce a una sola unidad –o sea a una única variable- se dice que “ha degenerado”. En efecto, en ambos casos se termina renunciando a la posibilidad de la **comparación**, procedimiento intelectual que se encuentra en la base misma de la posibilidad del conocimiento. Como explica muy bien Galtung, existen varias razones por las que resulta conveniente evitar estas situaciones. Así, cabría preguntar “¿Por qué esta unidad y no otra?” Suponiendo en efecto que se pretendiera estudiar un sistema social. ¿Qué justificaría elegir un informante antes que otro? No podemos sostener la creencia en el informante “puro”; la mejor prueba de ello es que si optáramos por otra unidad de análisis podríamos llegar a conclusiones muy distintas sobre nuestro objeto.¹⁶ La posibilidad misma de constatar variaciones y diferencias, y de evitar caer en estereotipos, requiere el poder comparar entre sí varias UA. Por la misma razón, tampoco conviene trabajar con una sola variable: no existe la pregunta “pura”.¹⁷ Disponer de varias variables permite

comparar respuestas entre sí, y de este modo poder situar en un contexto la respuesta presentada en una variable, así como detectar patrones de respuesta.¹⁸

Finalmente, si el objetivo del conocimiento científico es establecer relaciones entre variables, su instrumento privilegiado habrá de ser la correlación –esto es, comparar valores en varias variables para un conjunto de UA. Y como se entenderá, no es posible el uso de la correlación sino a partir de un cierto número de unidades y contando por lo menos con dos variables.

La forma de la matriz de datos permite pensar con la mayor claridad la articulación entre las tres series de elementos (U_{ai} , V_j y R_{jk}) que concurren en la constitución del dato. Sin embargo, no siempre esta articulación se percibe nítidamente de modo inmediato. Un caso interesante se plantea cuando consideramos la misma unidad de análisis en distintos puntos del tiempo.

“Misiones” en 1991 presentará muy probablemente un valor distinto en la variable “*Nº de hogares con NBI*” que el anotado para 1980. En rigor, cabría inquirir: si “Misiones” en 1991 es la misma unidad de análisis que en 1980 o si se trata de dos unidades distintas; es obvio que en el fondo la respuesta a una pregunta de este tipo es materia de convención.

Interesa destacar en términos prácticos cómo se puede zanjar técnicamente esta cuestión. ¿Qué hacer cuando no se trabaja con varias UUAA sino con una sola, la que ha sido medida en el mismo conjunto de variables en distintos puntos del tiempo? Supongamos que contamos con un conjunto de datos para una UA, la Universidad Nacional de Misiones, medida en dos variables “*Nº de Nuevos inscriptos (NI)*” y “*Nº de Desertores*” en cinco años sucesivos. Podríamos, por supuesto, presentar los datos bajo la forma de una matriz que hubiera degenerado en un único renglón, como en este ejemplo:

UA	NI82	NI83	NI84	NI85	NI86	De82	De83	De84	De85	De86
UnaM	858	983	1349	2358	1996	387	561	947	1530	1248

Nada nos impide presentar nuestros datos bajo la forma de este vector- Procediendo de este modo, estamos midiendo nuestra única UA –la UnaM- en diez variables distintas: en efecto, NI82 es una variable, NI83 otra, etc. Claro que es ésta una manera impráctica de presentar este conjunto de datos; no sólo esta forma es conceptualmente oscura sino que, en especial, es impropia para ser trabajada mediante cualquier programa estadístico. Obsérvese que al disponer de una sola UA, no es posible el cálculo de ninguna correlación.

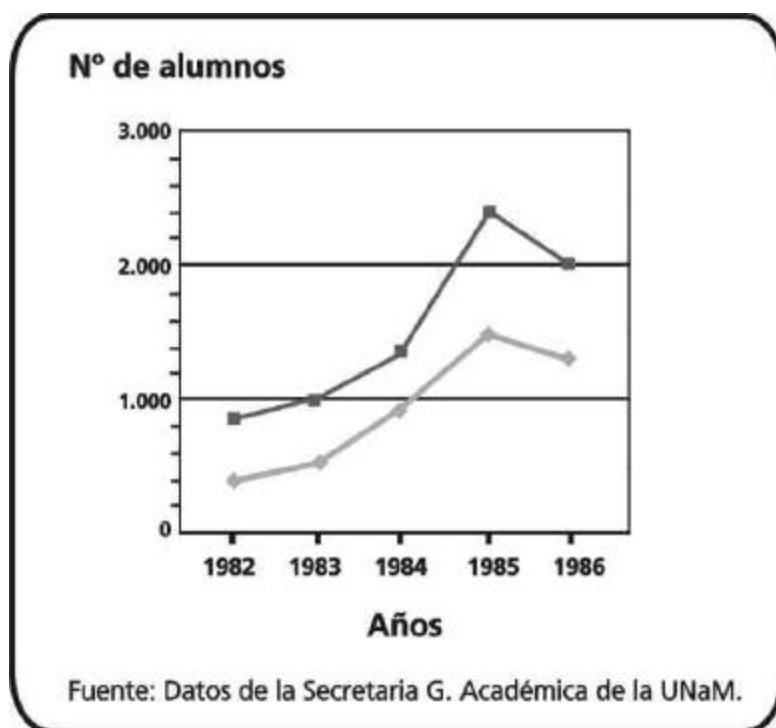
Existe una alternativa mucho más eficiente: y es tratar a nuestra UA como si fueran otras tantas UA, como medidas en diferentes puntos en el tiempo que disponemos. Es decir, asumimos que una UA es idéntica a sí misma, sólo en un determinado punto del continuo temporal. Debemos entonces presuponer que $U_{at} = U_{at+1}$. De hecho, así es cómo se resuelve técnicamente la cuestión. UNaM82 y UNaM83 funcionarán como dos unidades de análisis distintas. En la matriz que generemos, cada renglón corresponderá entonces a un año diferente.

UA	NI	De
1982	858	387
1983	983	561
1984	1349	387
1985	2358	1530
1986	1996	1242

Los mismos datos se han reestructurado asumiendo convencionalmente que corresponden a cinco UUAA medidas en dos variables. Bajo esta nueva presentación, salta a la vista la relación existente entre las dos variables. El crecimiento del número de desertores nuestra una clara asociación con el incremento en el número de nuevos inscriptos.

Ahora bien, suponiendo que deseemos graficar estas dos series, la siguiente figura nos permite reconsiderar la cuestión desde un tercer punto de vista. Conceptualmente, se puede pensar ahora que estamos en presencia de una sola unidad de análisis –la UnaM- medida en tres variables: el año –representado en el eje horizontal-, su número de nuevos inscriptos y su número de desertores –ambas medidas en el eje vertical.¹⁹

Figura IX. Evolución e los nuevos inscriptos y desertores en la UnaM, 1982-86



Ejemplo que debería terminar de convencernos acerca de este punto: no existe ninguna posibilidad de aplicar mecánicamente la estructura de la matriz a un conjunto de observaciones, sino que se requiere de todo un trabajo conceptual previo que es el que realmente produce –o “da”- los datos. En la sección siguiente, partiendo de la idea de “niveles de inclusión”, ahondaremos algo más en esta cuestión.

¹ Es esencial comprender que **todo** enunciado fáctico es, por definición, susceptible de ser reducido a estos tres elementos. No queremos decir con esto que siempre haya que proceder así (existen, por ejemplo, enunciados no-fácticos con sentido); ni tampoco que ésta sea la única ni la mejor manera de obrar en todos los contextos (a un psicoanalista este esquema no le aportaría nada para interpretar al discurso de su paciente). Sí, en cambio, sostenemos que no existen investigación empírica cuantitativa que no descansen en algún tipo de matriz de datos.

² Datos de Argentina, 1984.

³ Otra posibilidad de comparación es –como se verá luego- considerar la misma unidad de análisis en distintos puntos del tiempo.

⁴ Operacionalizar una proposición supone contar con una definición precisa de la unidad de análisis. En este caso, se podría definir como “electores” a todas las personas inscriptas para esa fecha en el padrón electoral de Posadas, por ejemplo.

⁵ Obsérvese que hemos adjudicado en forma arbitraria los códigos numéricos “1” y “2” para simbolizar los valores “masculino” y “femenino”.

⁶ Es evidente que en cuanto a esta segunda variable “*dirección del voto*” se podrían distinguir otros tantos valores como partidos se presentaron a dichas elecciones, pero dos valores son suficientes para traducir adecuadamente el significado de la proposición que no ocupa. Este ejemplo es ilustrativo de cómo cualquier clasificación politómica (de más de dos categorías o valores) puede reducirse a una **dicotomía** (una clasificación de dos valores).

⁷ En álgebra matricial, se define más genéricamente como “matriz” a “cualquier conjunto rectangular de números”. (Cf. Namocodiri, 1984:8)

⁸ Obsérvese que si la relación hubiera sido $3/5 > 2/5$, también se verificaría la hipótesis; en efecto, ésta se encuentra formulada con escasa precisión. Por otra parte, a los efectos pedagógicos prescindimos por el momento de toda consideración sobre la **significación** estadística del resultado (que sería nula, con una muestra de $n=10$)

⁹ Los valores de esta variable podrían corresponder a las 23 Provincias más Capital Federal y se codificarían de “01” a “24”.

¹⁰ Hemos modificado la simbología de Galtung. Así, hablamos de n (y no m) unidades de análisis, lo que presenta la ventaja de ser congruente con el uso habitual –tanto antes como después de Galtung– que se le da al símbolo “ n ” (o “ N ”) para designar al número de casos (o de UUAA) seleccionados para una muestra o de elementos existentes en el universo. En aras a una mayor simplicidad de exposición, renunciamos también diferenciar “variable” de “estímulo” (y “objeto”), “unidad de análisis” de “sujeto”, “valor” de “respuesta”.

¹¹ Es también la forma más práctica de disponer los datos para su tabulación manual.

¹² Cf. 1966:I, 4-6.

¹³ En el caso en que al encuestado se le permite marcar varias respuestas, se le está presentando bajo una misma pregunta varias variables, una para cada categoría de respuesta, de tal manera que se encuentra ante dos valores posibles para cada variable: aceptación o rechazo.

¹⁴ “Como regla general de tipo práctico, puede decirse que un 10% es el máximo absoluto de celdas vacías admisibles en cualquier columna o cualquier fila de M y que un 5% es un máximo más aconsejable” (Cf. Galtung, 1966: I,6).

¹⁵ Cf. 1966: I,9.

¹⁶ “En Química o Física a menudo parece haber sido resuelto el problema de encontrar el caso puro. Si un químico desea comprobar una proposición acerca del sulfuro, puede tomar una cantidad cualquiera de sulfuro químicamente puro (siempre que su forma cristalina sea irrelevante para el experimento) y tratarlo como si fuera un puro y verdadero representante del sulfuro, S ” dice Galtung (1966:I, 10). Pero éste no es el caso en las ciencias sociales.

¹⁷ Constatación que se encuentra a la base misma de la necesidad de elaborar índices –esto es, variables complejas– para representar dimensiones complejas de una clase de fenómenos (Cf. *Infra*, cap. V)

¹⁸ El mismo Galtung señala excepciones a esta necesidad de contar con varias variables: “Cuando se busca la simple información acerca del contexto, o cuando se investiga una dimensión de actitud que ha ocurrido muy a menudo en el debate diario, quien investiga a través de una encuesta parece justificado al limitarse a una sola pregunta. Las encuestas de Gallup caen en una de estas categorías o en ambas, lo que las hace metodológicamente justificables desde este punto de vista” (1966: I, 9)

¹⁹ Desarrollando una idea de Cattell de 1952, Rogers y Kincaid proponen la idea del “cubo de datos (*Data-Cube*) una matriz tridimensional construida sobre la base de (1) Unidades de Análisis (individuos o relaciones), (2) Variables y (3) Tiempo (Cf. 1981: 79-81).

NOCIONES DE MUESTREO

Cualquier investigación empírica versará sobre sistemas reales situados en el espacio y en el tiempo. El estudio de un objeto determinado es algo que puede demandar la definición de uno o varios universos. El fin que se persiga puede ser eminentemente práctico –describir qué piensan los obreros de una fábrica o cuáles son las intenciones de voto de los habitantes de una provincia, u obtener una estimación de la tasa de desempleo en una ciudad, etc.- o bien responder a una motivación más estrictamente científica –como cuando se trata de fundar hipótesis de un mayor grado de generalidad.

La generalidad es algo que se puede entender por lo menos en dos sentidos. Zetterberg distingue entre proposiciones teóricas y comunes basándose en cuál sea su “valor informativo”¹. Si se dice “A mayor educación, mayor ingreso monetario”, ésta puede ser una hipótesis de un nivel de generalidad muy elevado en el sentido de que, potencialmente, es aplicable a cualquier sociedad en la que se reconozca la existencia de algo llamado “educación” y en la que funcione alguna forma de moneda para regular el acceso de sus miembros a bienes y servicios. Sin embargo, no es ésta una hipótesis de alto valor informativo en cuanto a la variedad de fenómenos de los que da cuenta.

El problema del alcance de una hipótesis es entonces doble: por una parte, tiene que ver con sus referentes concretos –con la cantidad de sistemas o UUAA comprendidos dentro de su dominio-; por la otra, con el grado de abstracción de sus conceptos de propiedad respecto a conceptos más específicos –con la cantidad, por lo tanto, de conceptos de propiedad que subsumen. Así, cuando se dice “Toda persona tiende a realizar aquellas acciones que ayudan a mantener invariables las evaluaciones que recibe de sus iguales” se trata de una proposición de un alto nivel de generalidad en este segundo sentido. En efecto, como lo señala Zetterberg, de ésta pueden deducirse varias proposiciones más específicas; por ejemplo, los políticos intentarán mantener su popularidad, los investigadores elegirán temas que les permitan conservar su prestigio como miembros de una comunidad científica, etc. Un concepto tan abstracto como el de “evaluaciones” estará subsumiendo varios conceptos de menor nivel (“aprobación”, “estima”, “rango”); por lo que la mera sustitución de dicho concepto en la proposición más abstracta, permitirá generar toda una serie de hipótesis específicas.

Si se admite que el desarrollo del conocimiento científico supone tender hacia la formulación de hipótesis “generales abstractas”, se entenderá que tanto el grado de generalidad como el de “abstracción” son relevantes en cuanto a determinar el alcance de una hipótesis cualquiera. Así como era posible partiendo de la idea de “niveles de inclusión” determinar superunidades (“colectivos”), así también podría aplicarse la idea de inclusión a los conceptos de propiedad, obteniendo superconceptos.

El problema clásico de la inducción deviene de la imposibilidad lógica de sustentar la verdad de hipótesis generales a partir de un número de observaciones limitado y siempre finito. Ahora bien esta limitación del “número de observaciones” se refiere tanto a las UUAA como a las variables: ni estudiamos todos los elementos de un conjunto de UUAA, ni de todas las UUAA observadas consideramos todas las propiedades. Desde este punto de vista la técnica del muestreo podría considerarse como una alternativa estadística para superar el problema de la inducción. Empero esta técnica es más apta para resolver el problema en cuanto al número de las UUAA. Las razones por las que no se puede aplicar con el mismo éxito a las variables son varias; mencionemos simplemente la que deviene de la dificultad en delimitar un universo de variables. Por ende, nos

limitaremos a plantear el problema del muestreo en lo referido al problema de la selección de un cierto número de UUAA. Abordaremos luego la cuestión del tamaño de la muestra, y finalmente describiremos los diferentes tipos de muestras que se utilizan más corrientemente.

1. ELEMENTOS BÁSICOS DE MUESTREO

Ya hemos introducido el concepto de “universo” para referirnos a un conjunto de UUAA definidas por presentar valores constantes en un conjunto de variables. Indiferentemente puede utilizarse este término o el de “**población**” para denotar “el conjunto de todos los casos que concuerdan con una serie determinada de especificaciones”². Así, cuando se habla comúnmente de “la población de la Argentina”, se está haciendo referencia a un conjunto de UUAA –“personas”- que en un momento dado (por ejemplo, el 30 de setiembre de 1980), en la variable “país de residencia” presentan el valor “Argentina”. En la jerga estadística, una población se compone de “elementos”: así, cada persona habitante de la Argentina es un **elemento** de dicha población. Contrariamente al uso corriente del término “población”, los elementos no tienen por qué ser personas. Podrá tratarse también de grupos, acciones, productos o cualquier tipo de objetos o sistemas que se convenga en definir como tales: cada uno de los avisos transmitidos por un canal de TV durante una semana, de los informes incluidos en un archivo, de los desplazamientos turísticos realizados por los habitantes de región durante un período dado, de las relaciones existentes entre los miembros de una comunidad, etc., pueden ser considerados elementos de las correspondientes poblaciones. En cualquier caso una población incluye **todos** los elementos que satisfacen las propiedades que la definen: todos los obreros de una fábrica, todos los departamentos o partidos en que se dividen las provincias argentinas, todos los municipios de Misiones, todas las noticias sobre suicidios de jubilados publicadas en diarios de Buenos Aires durante 1992, etc.

Desde este punto de vista uno de los primeros pasos en cualquier investigación empírica consiste en la determinación del universo de la investigación. Así, por ejemplo, en un trabajo realizado para la Entidad Binacional Yacypetá,³ el universo del estudio fue definido como constituido por unidades domésticas que presentaran las siguientes características: pertenecer a la ciudad de Posadas, estar ubicadas por debajo de la cota 85, dentro de la Zona 1 de relocalización y cuyos miembros no fueran propietarios del terreno en que se asentaban.

Así como distinguíamos sub-universos dentro de un universo, así es posible definir poblaciones que se encuentran incluidas dentro de poblaciones mayores. En el contexto de la teoría del muestreo tales sub-poblaciones reciben el nombre de “estratos de la población”, o simplemente “estratos”. Un **estrato** se define a partir de una o más variables que permiten dividir a una población en sus conjuntos mutuamente excluyentes. Dentro de la población de los electores de Posadas la variable “sexo” permitirá definir el estrato de los varones y el de las mujeres; dentro del universo de los establecimientos industriales de la Argentina se puede distinguir estratos de acuerdo con la variable “número de empleados”: los que ocupan hasta 5 empleados, de 6 a 50 empleados, y de más de 50 empleados, por ejemplo. Combinando las variables “sexo” y “condición de alfabetismo”, la población de los votantes de Posadas el 14 de mayo de 1989, podría dividirse en cuatro estratos, varones alfabetos, varones analfabetos, mujeres alfabetas y mujeres analfabetas, etc.

Supongamos que deseamos conocer qué proporción de una población presenta un determinado valor en una variable dada. Si la población es pequeña, ello no ofrece mayores problemas. Por ejemplo, para conocer el porcentaje de mujeres entre los

alumnos presentes en un aula, se puede rápidamente contar cuántos alumnos hay y cuántos son de sexo femenino. Se denomina **censo** a este procedimiento, que consiste en “un recuento de todos los elementos en una población y/o una especificación de las distribuciones de sus características, basados en la información obtenida para cada uno de los elementos”.⁴

Si por lo contrario, la población que nos interesa es de gran tamaño, por ejemplo, los habitantes de Buenos Aires en 1993, podría resultar sumamente trabajoso obtener información sobre todos sus elementos. En este caso convendrá utilizar una **muestra**, vale decir un subconjunto de elementos de una población seleccionados para averiguar algo sobre el conjunto total de elementos que constituye esa población.⁵ Cuando se utiliza una muestra, se lo hace en base a la creencia de que la exactitud de los resultados obtenidos de ese modo es lo más próxima posible a la que se hubiera obtenido realizando un censo desprovisto de todo tipo de error sobre la población. Por supuestos, tal creencia constituye una hipótesis cuya verdad sólo sería demostrable realizando paralelamente un censo sobre la misma población.⁶ Por una parte existen los **parámetros** poblacionales, es decir los valores “verdaderos” que caracterizan las distribuciones de variables en la población: por ejemplo, la media de ingresos de todos los jefes de unidades domésticas de Buenos Aires, o el porcentaje de todos los electores con intenciones de votar por el Justicialismo en las próximas elecciones; dichos parámetros son los valores que se obtendrían en una medición de todos los elementos desprovista de cualquier tipo de error. Por la otra, se obtienen **valores muestrales**, simples **estimadores** de aquellos parámetros. Lo cierto es que nunca resulta posible asegurar la coincidencia entre los valores muestrales y los parámetros; a lo sumo, la forma de selección de la muestra podrá maximizar nuestra creencia en dicha coincidencia.

El problema del muestreo consiste entonces en cómo seleccionar una muestra de modo tal de obtener la máxima aproximación a los parámetros poblacionales compatible con las restricciones de costo y de tiempo existentes. En efecto, si se dispusiera de una cantidad ilimitada de recursos, lo que cabría sería recurrir a un censo. Pero lo cierto es que las más de las veces no es ésta la situación en que se encuentra el investigador.

A menudo se habla de la “representatividad” de la muestra. La idea de representatividad tiene que ver con la posibilidad de que la muestra sea una réplica adecuada de la población en lo que hace a las variables relevantes. Al respecto, convendrá tener en cuenta las consideraciones siguientes.

1. Una muestra en particular no es representativa en abstracto; una muestra que sea representativa para determinados propósitos, puede no serlo para otros: no existe la muestra representativa en sí, para cualquier propósito.
2. Si bien existen procedimientos para evaluar la bondad de una muestra, no es la muestra en sí la que es representativa; más bien es nuestra creencia en su representatividad la que va a depender del plan de muestreo utilizado para seleccionarla. Por “**plan de muestreo**” se entiende el diseño de un conjunto ordenado de operaciones que conduce a la selección de un número determinado de unidades de análisis a partir de un universo determinado.

2. EL TAMAÑO DE LA MUESTRA

Sin duda, determinar el tamaño adecuado para una muestra es una elección crucial, por todas las consecuencias negativas que puede producir una equivocación. No obstante ello, no es infrecuente que el tamaño de la muestra se determine en función de los recursos disponibles. Está claro que si la muestra es demasiado grande, se habrá realizado una inversión inútil, un derroche; en cambio, si resulta exigua, ocurrirá que no servirá a los propósitos de la investigación.⁷ Lamentablemente, y contrariamente a

opiniones corrientes, no existe ninguna regla áurea del tipo “tomar un 10% de la población” para asegurarse el éxito. En algunos casos 10% resultará lastimosamente insuficiente, mientras que en otros será un total exceso.

Atendiendo a una distinción de Galtung, todas las muestras pueden clasificarse en dos grandes categorías según el tipo de hipótesis que se pretende poner a prueba. En efecto puede tratarse tanto de **hipótesis de generalización** –cuando a partir de los datos muestrales se pretende inferir, por ejemplo, el valor de un parámetro en una población dada-, o bien de **hipótesis sustantivas** –cuando se desea comprobar la existencia de determinada relación entre ciertas variables. En el primer caso los criterios pertinentes a tener en cuenta serán los habituales para cualquier tipo de inferencia estadística; en el segundo, pueden resumirse en unas breves reglas muy sencillas de comprender.

2.1. Muestras para verificar hipótesis de generalización.

Existen procedimientos estadísticos que nos permiten estimar la probabilidad de que un determinado valor muestral no difiera sustancialmente del parámetro que se hubiera obtenido de haber realizado un censo. Así, si se desea averiguar qué porcentaje de los elementos de una población presenta el atributo X, es posible adoptar un plan de muestreo tal que nos garantice, por ejemplo, con un **nivel de confianza** del 95% que el parámetro no se apartará más de un 2% (**margen de error**) del valor muestral que obtengamos. Si requiriéramos mayor seguridad y precisión elegiríamos un nivel de confianza mayor (99%, por ejemplo) y un margen de error menor (1%); empero, cuanto más alto sea nuestro nivel de aspiraciones, mayor deberá ser el tamaño de la muestra, lo que supondrá un costo más elevado, puesto que deberemos seleccionar y recolectar datos sobre mayor cantidad de unidades. A los efectos prácticos un plan de muestreo con un nivel de confianza del 99% significa que, siguiendo ese plan de muestreo, 99 veces de cada 100 no nos equivocaremos en nuestra estimación; empero, no existe nunca garantía alguna de que la muestra particular que haya resultado seleccionada no pertenezca a aquel conjunto del 1% de las muestras que nos conducirían a error.

Tabla A: Tamaños de muestras al azar simple requeridos para niveles de confianza del 99,7% y del 95,5%, según valores presumibles de p y q y límites de error.

Nivel de confianza de 99.7% (3 Ó)					
Límites de error (+/- %)	Valores presumibles de p y q en %				
	10/90	20/80	30/70	40/60	50/50
1,0	8.100	14.400	18.900	21.600	22.500
2,0	2.025	3.600	4.825	5.400	5.627
3,0	900	1.600	2.100	2.400	2.500
4,0	506	900	1.181	1.350	1.406
5,0	324	576	756	864	900
10,0	81	144	189	216	225
20,0	20	36	47	54	56

Nivel de confianza de 95,5% (2 Ó)					
Límites de error (+/- %)	Valores presumibles de p y q en %				
	10/90	20/80	30/70	40/60	50/50
1,0	3.600	6.400	8.400	9.600	10.000
2,0	900	1.600	2.100	2.400	2.500
3,0	400	711	933	1.067	1.111
4,0	225	400	525	600	625
5,0	144	256	336	384	400
10,0	36	64	83	96	100
20,0	9	16	21	24	25

Fuente: Sierra Bravo, 1979: 181–182.

La Tabla A ilustra cómo el tamaño requerido para una muestra es una función directa de tres factores.

1. El **nivel de confianza** requerido: la tabla está dividida en dos mitades. En las celdas del sector superior se presentan los tamaños de muestra requeridos para un nivel de confianza del 99,7%; en las del sector inferior los tamaños demandados por un menor nivel: 95,5%. Puede observarse cómo en el sector inferior los números son menores a sus análogos del sector superior.
2. El **margen de error**: en la primera columna se indican distintos márgenes de error para cada una de las hileras de la tabla. El menor límites de error incluido es de +/- 1%; vale decir que si con ese límite producimos una estimación de que un 40% de las unidades observadas en la muestra poseen el atributo, podremos inferir que la proporción en la población estará ubicada entre el 39 y el 41%. En ambos sectores de la tabla de abajo hacia arriba crecen los tamaños requeridos: cuanto menor sea el margen de error que deseemos, mayor deberá ser el tamaño de la muestra.
3. La **variabilidad** del atributo investigado: cada columna corresponde a diferentes distribuciones del atributo en el universo, que van desde la distribución 10/90 (10% poseen el atributo, 90% no lo poseen) hasta 50/50 (una mitad posee el atributo, la otra no). En cada sector los tamaños aumentan a medida que nos desplazamos de izquierda a derecha: cuanto mayor es la variabilidad de la característica,⁸ mayor es el tamaño de muestra necesario. Así por ejemplo, mientras que para un nivel de confianza del 95,5 y un margen de error de +/- 2%, 900 casos serían suficientes si el atributo se distribuyera en las proporciones 10/90, en el caso de que la distribución en el universo fuera 50/50 la muestra requerida sería de 2.500 unidades.⁹ Este último factor supone la necesidad de anticiparse a la distribución del atributo en el universo; y de no existir elementos que permitan sustentar otra alternativa, convendrá conservadoramente suponer la distribución más desfavorable: 50/50.

En síntesis, cuanto mayores sean nuestras exigencias respecto al grado de confiabilidad y de precisión de nuestros resultados, mayor habrá de ser el tamaño de la muestra. La tabla permite también apreciar cómo juegan aquí rendimientos decrecientes. Para un atributo cuya distribución en el universo es de 20/80, 256 casos alcanzan para determinar el valor del parámetro con un nivel de 95,5 y un margen de error de +/- 5%; si se lleva la muestra a un número de 711 unidades, el margen de error se reduce en dos puntos, pasando a ser de +/- 3%. Pero, para obtener una nueva reducción de dos puntos y alcanzar un margen de +/- 1%, el aumento en el tamaño deberá ser mucho mayor, requiriéndose ya 6.400 casos.

2.2. Muestras para someter a prueba hipótesis sustantivas.

Si el objetivo de nuestra investigación es primordialmente analítico antes que descriptivo, no estaremos tan interesados en la generalización como en la simple comprobación de la existencia de relaciones específicas entre las variables. Desde este punto de vista, el tamaño no se basará en los requerimientos demandados para producir inferencias estadísticas, sino que deberá ser tal que nos permita determinar la existencia de relaciones para el conjunto de las unidades incluidas en la muestra.

Si por ejemplo queremos estudiar la relación entre dos variables dicotómicas, ello demandará producir una tabla de contingencia en la que las unidades se clasifiquen simultáneamente en ambas variables.¹⁰ Una tabla tal consta de cuatro celdas, y para aplicar una simple diferencia porcentual difícilmente nos contentemos con un promedio de casos por celda inferior a 10. Vale decir que **con menos de 40 casos no es posible analizar ninguna relación entre variables**; incluso, para que la tendencia sea clara, preferiremos sin duda contar con un promedio de 20 casos por celda: aún así el simple cambio de un caso de una columna o de una hilera a otra producirá una diferencia porcentual del 2,5%.

En esta perspectiva, las preguntas a formularse para determinar el tamaño de la muestra tienen que ver con el número de variables que se pretende analizar simultáneamente y con el número de valores de cada una de esas variables. Galtung presenta la siguiente tabla:

Tabla B: Número mínimo de unidades de análisis para un promedio de 10 casos por celda (20 casos entre paréntesis)

Nº de variables por tabla	Nº de valores por variable		
	2	3	4
2	40 (80)	90 (180)	160 (320)
3	80 (160)	270 (540)	640 (1.280)
4	160 (320)	810 (1.620)	2.560 (5.120)

Así, por ejemplo, para estudiar la relación entre tres variables tricotómicas se requerirá un mínimo de 270 casos para contar con un promedio de 10 observaciones por celda (ó 540, si pretendemos tener 20 casos en cada combinación). Obviamente no es indispensable que todas las variables tengan el mismo número de valores, ni que deban limitarse a cuatro valores, lo que puede dar lugar a otros tamaños de muestra. Lo más práctico es entonces multiplicar el número de casos por celda por el número de valores de cada una de las variables que se desee analizar simultáneamente; si se quisiera investigar cómo se relacionan entre sí dos variables dicotómicas y una tercera de cinco valores, el número mínimo de casos requerido para contar con un promedio de 10 observaciones por celda sería $(10) \times 2 \times 2 \times 5 = 200$. Por supuesto, en cualquier investigación es posible pretender analizar la relación entre varios conjuntos de variables, por lo que la pregunta que debe responderse para determinar el tamaño de la muestra es cuál es la tabla más grande que se pretende generar.

En las muestras para hipótesis sustantivas es posible de todos modos aplicar determinados test de hipótesis como la de Student aplicada a la diferencia de medias o de proporciones, o la prueba de χ^2 . En estos casos, puede interpretarse que la hipótesis sustantiva pasa a funcionar también como una hipótesis de generalización, aunque no ya referida a una población finita y concretamente determinada, sino con relación a un universo teórico, sin límites definidos, que abarque las unidades de todo tiempo y lugar a las que sea posible aplicar las variables.

3. PRINCIPALES TIPOS DE MUESTRAS

No es nuestro propósito reseñar aquí todos los tipos posibles de muestras. Tampoco nos proponemos abordar este tema en una perspectiva estadística; ante cualquier investigación de envergadura, convendrá recurrir a algún especialista en la cuestión.¹¹ Nuestra intención es sólo describir las características de algunos diseños muestrales desde un punto de vista conceptual, para exponer algunos problemas típicos que se plantean con relación al uso de esta técnica. Desde el punto de vista de la teoría del muestreo, existen solamente dos grandes categorías de muestras. Por un lado están las muestras probabilísticas –o “al azar”–, que se caracterizan porque permiten especificar la probabilidad que tiene cada elemento de la población de resultar incluido en la muestra. El caso más sencillo es cuando todos los elementos tienen la misma probabilidad de integrar la muestra, pero no es éste el único tipo de muestreo de probabilidad. Por otra parte, hay muestras no-probabilísticas, en las que no existe forma de determinar la probabilidad que tiene cada elemento de ser seleccionado, y ni siquiera la seguridad de que cada elemento tenga alguna probabilidad de ser seleccionado.

Esta distinción es fundamental, puesto que solamente los planes de muestreo probabilísticos pueden ser considerados representativos en un sentido estadístico. El muestreo probabilístico es el único que permite basarse en la teoría de las probabilidades para estimar estadísticamente el grado en que los valores muestrales pueden tender a diferir de los parámetros; una muestra probabilística permite especificar el tamaño de muestra requerido para un nivel de confiabilidad y un margen de error determinados.

3.1. Muestras no-probabilísticas

En el muestreo de no-probabilidad no hay modo alguno de evaluar estadísticamente los resultados obtenidos a partir de la muestra. No obstante ello, algunas muestras de este tipo se utilizan mucho, en razón de las ventajas comparativas que exhiben en cuanto a su comodidad y bajo costo.¹²

Muestras accidentales (o “casuales”): consisten simplemente en tomar los casos que “caen bajo la mano” continuando el proceso hasta que se alcanza un cierto tamaño de la muestra. Por ejemplo, se incluirán en la muestra las primeras 100 personas que pasen por una esquina y que sean deseosas de ser entrevistadas. Si el universo es la población de una ciudad, es claro que no será ésta una muestra representativa: no todos los elementos de dicha población tienen la misma oportunidad de pasar por dicha esquina, ni menos aún coincidirán en la probabilidad de pasar por allí en el momento en que se realicen las entrevistas. Pero en este caso, no existe modo alguno de evaluar los sesgos introducidos por el modo de elección de las unidades; sólo cabrá esperar que la equivocación no sea demasiado grande.

Muestras por cuotas: en el muestreo por cuotas se busca garantizar la selección de elementos pertenecientes a los diferentes estratos que conforman la población y que dichos elementos puedan ser tenidos en cuenta en las mismas proporciones en que ocurren en esa población. En todos los casos, la técnica supone sostener alguna hipótesis acerca de cuáles son las variables relevantes, las que serán utilizadas para dividir la muestra en estratos. Si, por ejemplo, se quiere conocer las intenciones de voto de una población, y se piensa que éstas variarán en función del sexo, convendrá estratificar la muestra en dos estratos: “masculino y femenino”. Así, a cada entrevistador se le dará la consigna de cumplir con una “cuota” de entrevistados pertenecientes a cada estrato: tantos varones y tantas mujeres. El muestreo por cuotas requiere anticiparse a las diferencias que puedan presentar los elementos de los distintos estratos respecto al valor

que exhibirán en la variable a investigar, de modo tal de seleccionar una muestra que “represente” adecuadamente a la población, esto es que funcione como una réplica de la población en cuanto a las variables que se juzgan relevantes. En lo que respecta al estudio de preferencias, opiniones, actitudes, etc., se conoce que variables tales como el sexo, la edad, o el nivel educativo tienden a producir diferencias significativas en esas características.

Ahora bien, puede suceder que la proporción de elementos de un estrato en la muestra no sea idéntica a la existente en la población. Imaginemos que se encuesta a 100 personas obteniendo una cierta distribución¹³ por nivel educativo, la que podemos comparar con la existente en el universo:

Tabla 3.1: comparación de la distribución de la variable “nivel educativo” en la muestra y en la población

Nivel Educativo	% muestra	% población	Diferencia porcentual
Alto	30	15	+15
Medio	45	30	+15
Bajo	25	55	-30
Total	100	100	

La comparación muestra claramente que en la muestra los niveles alto y medio se encuentran sobrerrepresentados, mientras que el nivel bajo está subrepresentado con relación a la población.

Tabla 3.2.: Distribución en la muestra de la intención de voto según el nivel educativo.

Intención de voto	Nivel Educativo			
	Bajo	Medio	Alto	Total
Justicialista	14	21	10	45 (45%)
No justicialista	11	24	20	55 (55%)
Total	25	45	30	100 (100%)

Ahora bien, supongamos que se ha preguntado a los entrevistados por su intención de voto en las próximas elecciones para estimar qué porcentaje de votos obtendrá el Justicialismo, obteniendo estos resultados. Si se los toma sin tener en cuenta el peso distinto que tienen los diferentes niveles educativos en la muestra y en la población se concluye que hay un 45% de los entrevistados que piensan votar por el Justicialismo.

Pero puesto que el nivel bajo está subrepresentado en la muestra, y que la proporción de votos por el Justicialismo es sustancialmente mayor en ese estrato, es evidente que el porcentaje de votos con esa orientación debe ser sustancialmente mayor en la población,¹⁴ no hay inconveniente alguno en calcular cuál hubiera sido el porcentaje de votos total por el Justicialismo si los distintos estratos hubieran estado representados en la muestra en las proporciones correctas. Para ello se recurre a simples reglas de tres:

las frecuencias de la columna “bajo” se multiplican por 55/25, las de “medio” por 30/45 y las de “alto” por 15/30:

Tabla 3.3.: Distribución corregida de la intención de voto según nivel educativo

Intención de voto	Nivel educativo			
	Bajo	Medio	Alto	Total
Justicialista	31	14	5	50 (50%)
No justicialista	24	16	10	50 (50%)
Total	55	30	15	100 (100%)

El resultado es ahora sustancialmente distinto: el marginal inferior exhibe la misma distribución de los distintos niveles educativos que se da en la población, y el porcentaje resultante de votos por el Justicialismo se eleva al 50%.

Lo que se ha hecho mediante estas sencillas operaciones aritméticas es restablecer el “peso” de los distintos estratos en la muestra, **ponderando** los resultados de acuerdo con la distribución conocida de la variable en la población. Por lo tanto, no es una condición necesaria para el muestreo por cuotas que los distintos estratos se encuentren representados en la muestra en las mismas proporciones que en la población, sino que bastará con que se cuente con un número suficiente de casos en cada estrato como para poder estimar los valores para los estratos de la población. Se denominan “**muestras autoponderadas**” aquéllas en las cuales los estratos se encuentran representados en idénticas proporciones que en la población.

Las muestras por cuotas son mucho más eficaces que las muestras accidentales simples. De allí que su uso sea muy frecuente en sondeos electorales así como en la investigación de mercado en general. Sin embargo, el muestreo por cuotas sigue siendo en esencia un procedimiento accidental en el que cada estrato de la muestra es una muestra accidental del correspondiente estrato de la población.

En particular, la libertad otorgada a los entrevistadores puede dar lugar a **muestras sesgadas**. Éstos son proclives a encuestar a las personas que tengan más a mano –por ejemplo, a amigos o conocidos-, a descartar a las personas que no se encuentren en su domicilio cuando los visitan, a evitar realizar entrevistas en barrios apartados, etc. Ello puede redundar en un cierto **sesgo** de la muestra, en una desviación con respecto a determinadas características reales de la población que la torne poco representativa para un objetivo determinado. Por cierto, ello dependerá del grado en que estas desviaciones se encuentren asociadas a variables relevantes con respecto al propósito de la investigación. Así puede darse que estén subrepresentadas en la muestra las personas que habitan en barrios apartados y que este efecto carezca de importancia en sí, pero que se produzca por esta vía una subrepresentación de las personas de bajo nivel económico-social debido a que éstas tienden a residir en barrios alejados del centro urbano, por ejemplo. Si se pudiera garantizar que no se ha producido ningún efecto de esta naturaleza en la selección de las unidades, el problema estaría resuelto. Pero como se trata de una muestra accidental no existe ningún fundamento que nos permita asumir esa hipótesis.

3.2. Muestras probabilísticas

En la investigación científica, el azar no es algo que se pueda presuponer como naturalmente dado: antes bien, se busca garantizar el azar, lo cual requiere recurrir a procedimientos con frecuencia caros y laboriosos. El azar científico se persigue de manera sistemática, es algo que se construye. Así, en las muestras probabilísticas –o

“aleatorias”- se requerirá que todos los elementos de una población tengan una probabilidad **conocida** de ser seleccionados.

Muestra al azar simple: es el ejemplo más sencillo de un procedimiento de muestreo probabilístico; pero aún más que por su posibilidad de ser aplicado directamente, su interés consiste en que se encuentra a la base de todos los otros tipos de muestras aleatorias. En la muestra simple al azar, todos los elementos de la población tienen la **misma** probabilidad de resultar seleccionados, y todas las combinaciones de elementos para un tamaño dado de la muestra presentan también la misma probabilidad de selección. Así, si nuestro universo está constituido por 100 elementos, cada uno de éstos tendría la misma probabilidad de ser incluido en la muestra: $p = 1/100$.

La probabilidad de una muestra de un tamaño dado de ser extraída, por su parte es igual a:

$$\frac{1}{\frac{N!}{N! (N-n)!}}$$

En la que: N = tamaño de la población
N = tamaño de la muestra
! = factorial

Así, para cualquier muestra de tamaño 10, extraída de una población de 100 elementos, la probabilidad de ser seleccionada será:¹⁵

$$p = \frac{1}{\frac{100!}{10! (100-10)!}} = \frac{1}{124.303\,388.140}$$

El requisito imprescindible para seleccionar una muestra al azar simple es disponer de un listado de todos los elementos que componen la población. A cada una de las unidades se le asigna un número desde 1 hasta N. Luego, recurriendo a una tabla de números al azar, a un bolillero, a una computadora, o a cualquier otro procedimiento que garantice la producción de una serie de números al azar, se extraen en la cantidad necesaria los números correspondientes a las unidades que integrarán la muestra. Así, si se tiene una población de $N = 1.000$ y se desea seleccionar una muestra de 100 unidades, se sortearán 100 números comprendidos entre el 1 y el 1.000, tomando los últimos tres dígitos de los números generados.

En la práctica, el procedimiento del muestreo al azar simple se torna por demás engorroso al trabajar con poblaciones grandes. Si además, por determinadas restricciones el número de casos en la muestra es pequeño, la muestra al azar simple puede conducir a una representación absolutamente insuficiente de los estratos de menor peso relativo en la población. Es por ello que en estas condiciones se suele optar por otros diseños muestrales libres de esas desventajas. No obstante, la importancia teórica de la muestra al azar simple es fundamental, dado que todos los otros diseños probabilísticos remiten a este modelo, y a que en todos esos otros diseños el muestreo al azar simple es utilizado al menos en alguna de sus etapas.

Muestra sistemática: se trata de un procedimiento de muestreo que, a la vez que es susceptible de simplificar notablemente la selección de las unidades, a todos los efectos prácticos presentará las mismas virtudes que el muestreo al azar simple. Aquí también se requiere como condición indispensable contar con el listado de todas las unidades, las que deberán ser numeradas correlativamente. Conociendo el número de

elementos que componen la población, y habiendo determinado el número de casos necesario para la muestra, se puede establecer la fracción de muestreo dividiendo el primer número por el segundo. Luego se selecciona un solo número al azar, que corresponderá a la primera unidad seleccionada; a este número se le suma la fracción de muestreo y se obtiene la segunda unidad; luego se suma nuevamente la fracción obteniendo la unidad siguiente, y se procede así sucesivamente hasta completar la muestra.

Supongamos que queremos extraer una muestra de 100 estudiantes de un universo constituido por una Facultad de 20.000 alumnos. Nuestra fracción de muestreo será: $20.000/100 = 200$. Seleccionamos un número al azar: el 4.387. Nuestra muestra quedará integrada por los estudiantes que lleven los números 4.387, 4.587, 4.787, 4.987, etc. Debe quedar claro que en este diseño ni todas las unidades ni todas las combinaciones de unidades tienen las mismas probabilidades de resultar seleccionadas. Así, en nuestro ejemplo, una muestra que incluya las unidades 4.387 y 4.388 tiene una probabilidad cero de ser seleccionada (dada nuestra fracción de muestreo de 200).

La condición para poder utilizar este procedimiento es que la numeración de las unidades en la lista no responda a ningún orden en particular, o por lo menos que de existir un orden éste no se base en ninguna propiedad susceptible de encontrarse relacionada con las variables que se desea estudiar. En determinadas circunstancias, sin embargo, la existencia de un orden puede resultar ventajoso pudiendo ponerse al servicio de la investigación. Imaginemos que queríamos estudiar las opiniones de los estudiantes de aquella Facultad respecto a cuestiones relacionadas con su organización académica. Como hipótesis, no resultaría irrazonable pensar que las opiniones podrían variar según la antigüedad de los alumnos en la institución. Ahora bien, si la numeración de la lista es tal que los números más bajos corresponden a los estudiantes más antiguos y los más altos a los más recientes, un procedimiento de selección de la muestra como el expuesto garantizaría la representación en la muestra de estudiantes con diferentes grados de antigüedad en la Facultad.

Muestra estratificada: siempre que hay razones para pensar que determinadas características de las UUA pueden encontrarse fuertemente relacionadas con las variables que tenemos interés en investigar, convendrá asegurar la representación en la muestra de los diferentes valores que pueden asumir las unidades en dichas características. Esto es, en base a determinadas variables que por hipótesis se asumen como pertinentes se divide la población en estratos. En el diseño más simple de muestra estratificada, la única diferencia con respecto a la muestra por cuotas es el hecho de que la selección de las unidades se realiza al azar. Por cierto, esta no es una diferencia menor, puesto que obtenemos una muestra apta para realizar cualquier tipo de inferencia estadística. Pero conceptualmente ambos tipos de muestra son muy similares: así como la muestra por cuotas consiste en una serie de muestras casuales tomadas cada una dentro de un estrato diferente, así una muestra estratificada se compone de varias muestras al azar simple seleccionadas dentro de otros tantos estratos.

Por supuesto, las mismas consideraciones que se hacían acerca de la ponderación de los estratos en la muestra por cuotas se aplican también al muestreo estratificado. Existen poderosas razones por las cuales puede resultar conveniente trabajar con un diseño de muestra que no sea autoponderado. En primer lugar, una muestra autoponderada puede llevar a seleccionar un número insuficiente de casos dentro de un estrato cuyo peso dentro de la población es pequeño. Así, por ejemplo, una muestra autoponderada estratificada por nivel económico social (NES), puede llevar a contar con un magro número de unidades dentro del estrato "Clase alta". Supongamos que la definición de "Clase alta" sea tal que comprenda a un 5% de la población.

Tomando una muestra de 400 casos, sólo se contaría con 20 elementos dentro de ese estrato; un número por cierto insuficiente para poder afirmar nada acerca de los integrantes de ese estrato. De ahí que pueda resultar interesante sobremuestrear voluntariamente ese estrato, de modo de contar, por ejemplo con 80 unidades con esa característica. Por supuesto, es factible hacer jugar esta consideración también en el muestreo por cuotas.

Pero en el caso del muestreo estratificado se agrega una razón más que hace a la eficiencia del procedimiento desde un punto de vista estadístico. En efecto, se ha visto que cuanto menor es la variabilidad de una característica en una población, menor será el número de casos necesario en la muestra para contar con una estimación adecuada del parámetro. Vale decir que para un número dado de observaciones muestrales, cuanto mayor sea la homogeneidad de la población, menor será el margen de error al estimar la distribución de la variable.

Dividir la población en estratos es útil en la medida en que los estratos sean internamente homogéneos y externamente heterogéneos en cuanto a las variables de nuestro interés. Si se cuenta con una estimación previa acerca de la variabilidad del parámetro dentro de cada estrato, entonces el modo más eficiente de muestrear¹⁶ es aquel en que el número de casos seleccionados sea proporcional a esa variabilidad. En la práctica, ello se resume en una consigna sencilla: sobremuestrear los estratos en que la variabilidad sea mayor, submuestrear los más homogéneos. Así, si en un sondeo sobre las intenciones de voto para las elecciones presidenciales se toman las provincias argentinas como estratos, se deberá sobre muestrear aquéllas en las que el resultado aparezca como más dudoso, por ejemplo. Lo expuesto debería bastar para comprender que, excepto por un pequeño ahorro de operaciones aritméticas, nada justifica muestrear a todos los estratos en la misma proporción y, por lo contrario, puede llegar a ser muy ventajoso no hacerlo.

Muestra por conglomerados y en etapas múltiples: si se quisiera hacer una muestra al azar simple —o sistemática, o aún estratificada— de la población de Buenos Aires, se tropezaría de inmediato con un obstáculo difícilmente salvable: la inexistencia de un listado de todos los habitantes de esta ciudad.¹⁷ Por cierto, nada impediría dedicar una etapa previa del muestreo a listar todos los elementos de esa población, a no ser pedestres consideraciones de costos.

Una solución podría ser recurrir a una muestra por **conglomerados** (*clusters*). Se trata de conglomerados de UUAA que funcionan como unidades de muestreo definidas espacialmente. Este tipo de diseño se utiliza en general para reducir los costos de la recolección de datos. En efecto, cuando el universo de la investigación resulta geográficamente extenso, es muy ventajoso, en términos económicos, tratar de concentrar la tarea en el espacio. Se reducen de este modo los costos que suponen el traslado del entrevistador de una a otra unidad.

Idealmente, los conglomerados son unidades de un mismo tamaño. Como bien observa Blalock (1966: 441), ésta es en cierto modo una estrategia opuesta a la del muestreo estratificado, aunque en ambos casos se divide a la población en grupos. Si en la muestra estratificada se seleccionan los casos **dentro** de cada estrato, asegurando así que todos los estratos estarán representados, en este otro tipo de muestra se selecciona **entre** los conglomerados; correlativamente, así como se busca que los estratos sean lo más homogéneos posible internamente, para los conglomerados cuanto más heterogéneos sean mejor será el resultado. Dividida toda la población en conglomerados, se selecciona al azar un cierto número de éstos dentro de los cuales se entrevistará a todas las unidades de análisis.

A menudo la técnica por conglomerados se integra en diseños de **muestra en etapas múltiples**, esto es, muestras que suponen la definición de unidades de muestreo en diferentes niveles. Así, si se quisiera medir la tasa de desempleo en la población de Buenos Aires con 14 o más años de edad, se podría estratificar la ciudad en zonas, y dentro de cada zona o estrato seleccionar al azar un cierto número de manzanas; la gran ventaja es que todas estas operaciones se realizarán en la comodidad de un escritorio si se cuenta con buena cartografía. Luego para cada una de las manzanas seleccionadas, se puede proceder en campo a un rápido relevamiento del número de unidades domésticas existentes. A partir de los listados para cada manzana, es posible seleccionar al azar unidades domésticas.¹⁸ Finalmente, en cada unidad doméstica el entrevistador podrá sortear por algún procedimiento adecuado la persona a encuestar. En este diseño multietápico, las manzanas funcionarían como unidades de muestreo de 1era. Etapa, las UDD serían unidades de muestreo de 2da. etapa, y las UUAA –personas- serían las unidades de muestreo de última etapa.

Sin duda, las muestras por conglomerados y en etapas múltiples son muy ventajosas desde el punto de vista económico, en el sentido que con igual cantidad de recursos es posible realizar mayor número de encuestas. Sin embargo, hay que tener en cuenta que con estas técnicas, contrariamente a la del muestreo estratificado, el grado de error aumenta: esto es para obtener el mismo grado de exactitud la muestra al azar simple se bastará con un menor número de casos.

Figura XIV: Resumen de los principales tipos de muestras

No-probabilísticas	Probabilísticas
Casual	Al azar simple
	Sistemática
	Estratificada
	Por conglomerados
	Por etapas múltiples
Por cuotas	

¹ Cf. Zetterberg; 1968: 66-67.

² Cf. Selltiz et al.; 1968:560.

³ “Estudio sobre los medios de subsistencia y la capacidad de pago de la población no propietaria de Posadas a relocalizar por Yacyretá” (Informe inédito producido por la FHCS, UnaM).

⁴ Cf. Selltiz et al.; 1968:561.

⁵ Nada impide combinar ambos procedimientos. De hecho el Censo Nacional de Población en la Argentina combina ahora un formulario corto que se aplica a todos los hogares (universo), y un formulario más extenso que sólo se aplica en una muestra de los hogares.

⁶ Aunque, obviamente, de ser este el caso ya no sería necesario recurrir a una muestra.

⁷ Existiría la posibilidad en ese caso de aumentar el número de unidades seleccionadas, pero ello entrañaría varias consecuencias negativas: a) el carácter de la muestra se vería desnaturalizado, en lo que hace a sus propiedades estadísticas; b) los costos de retornar al campo en busca de más datos pueden ser muy gravosos; c) se puede perder la comparabilidad entre los datos recolectados en dos períodos distintos, ya sea porque las variables objeto se hayan visto modificadas por el simple paso del tiempo, o porque el mismo proceso inicial de recolección haya conducido a una alteración en sus valores.

⁸ La variabilidad máxima es 50/50, vale decir cuando $p=q$ (siendo p la proporción de los que poseen el atributo y $q=1-p$). Intuitivamente, se comprende que si una población es absolutamente homogénea en cuanto a una característica dada, una muestra de cualquiera de los elementos de dicha población bastará para producir una estimación correcta del parámetro; en cambio si la población es heterogénea –se divide por mitades en varones y mujeres, por ejemplo– es claro que se necesitará un número mayor de observaciones para producir una estimación ajustada de la distribución de la característica en el universo.

⁹ Por ende, una muestra cuyo tamaño sea suficiente para estimar un determinado parámetro, podrá no serlo para otro parámetro, si la distribución de este último es menos favorable.

¹⁰ Cf. *Infra*: Cap. IV.

¹¹ O, al menos, remitirse a los capítulos pertinentes de los manuales de Blalock (1966), Padua (1979), Galtung (1966), Selltitz (1968), o Festinger y Katz (1975), o a textos específicos como el de Kalton (1987).

¹² Dada la naturaleza de este manual, nos concentramos en los tipos de muestras utilizados corrientemente con fines “cuantitativos”. En una perspectiva “cualitativa” –aunque también en la fase exploratoria de investigaciones que versen sobre un gran número de casos– cabe destacar la pertinencia de otros procedimientos de enorme valor heurístico. Así, “el **muestreo teórico** se efectúa para descubrir categorías y sus propiedades y para sugerir las interrelaciones dentro de una teoría. El muestreo estadístico se hace para obtener evidencia exacta sobre la distribución de la población entre categorías a ser usadas en descripciones o verificaciones. De este modo, en cada tipo de investigación la *muestra adecuada* que deberíamos buscar (como investigadores y lectores de investigaciones) es muy diferente.” (Cf. Glaser y Strauss; 1967:62.

¹³ Afortunadamente, se dispone con frecuencia de datos sobre la población a muestrear provenientes de otras fuentes. En este caso, los datos son imaginarios; pero si se tratara de muestrear la población de Posadas, por ejemplo, se podría recurrir fácilmente a los resultados de la Encuesta de Hogares para conocer la distribución de la variable en el universo.

¹⁴ Por ejemplo, se dispone de datos censales, o de la última onda de la Encuesta Permanente de Hogares realizada en la ciudad.

¹⁵ Cf. J. Padua (Ed.), 1979:66.

¹⁶ Vale decir, el que a igual cantidad de casos produzca mayor exactitud en la estimación del parámetro, o el que nos permita obtener al mínimo costo un resultado con un nivel de confianza y un margen de error dados.

¹⁷ Obsérvese que la utilización de la guía telefónica produciría una muestra sesgada, puesto que no todos los habitantes de Buenos Aires figuran en esa guía, y que la tenencia de teléfono, lejos de comportarse como una variable aleatoria, está asociada a otras características de las personas, como su nivel económico. A pesar de ello, las encuestas por vía telefónica son de uso frecuente (Cf. Lebart, 1992)

¹⁸ En la terminología del Censo de Población se habla tradicionalmente de “hogares” en vez de unidades domésticas.

TÉCNICAS ELEMENTALES DE ANÁLISIS

En la secuencia típica del proceso de investigación, el análisis de datos se inscribe dentro de las últimas etapas. Respondiendo a un esquema previamente establecido, las observaciones han sido realizadas, luego codificadas y tabuladas. El resultado es una serie de cuadros estadísticos a los que habrá que “leer” en función de los objetivos de la investigación, tratando de destacar lo esencial de la información en ellos contenida.

En este capítulo expondremos algunas de las técnicas más elementales que son de uso frecuente en la investigación social. Comenzaremos concentrándonos en la tabla de contingencia, puesto que éste es el modo en que tradicionalmente se presentan muchos datos en las investigaciones sociales. Acordaremos una atención privilegiada a la tabla de 2 x 2, en la que se plantean al nivel más simple los problemas lógicos del análisis. En este contexto, habré de referirme al uso de los porcentajes, al test de χ^2 y a algunos de los coeficientes de asociación más simples. Luego abordaremos brevemente otras formas de presentación de los datos como las distribuciones multivariantes conjuntas, lo que permitirá introducir algunas nociones sobre las relaciones entre variables pertenecientes a niveles más elevados de medición.

1. LA TABLA DE CONTINGENCIA Y EL USO DE LOS PORCENTAJES

Una tabla de contingencia es el resultado del **cruce** (o tabulación simultánea) de dos o más variables. Nos ocuparemos solamente de tablas bivariadas (o “bivariantes”), que también reciben los nombres de “clasificación cruzada” o “tabulación cruzada”. Esta forma de presentación de los datos es muy típica de la investigación en ciencias sociales, que se caracteriza por un uso predominante de variables (o atributos) definidas en los niveles de medición nominal y ordinal.¹ La tabla de contingencia consiste en un cierto número de celdas en las que, como resultado de un proceso de tabulación, realizando en forma manual, mecánica o electrónica,² se han volcado las frecuencias (número de casos) correspondientes a cada combinación de valores de varias variables.

Forma lógica de la tabla de 2 x 2

Para analizar la forma lógica de este tipo de tablas, consideraremos la estructura más sencilla, la llamada tabla de “2 x 2”, o sea de dos valores por dos valores, que resulta del cruce de dos dicotomías, los atributos “X” e “Y” en los que hemos clasificado un conjunto de unidades de análisis.

Figura XV: Forma lógica de la tabla de 2 x 2

		Atributo X		Total
		No	Sí	
Atributo Y	Sí	-XY	XY	Y
	No	-X-Y	X-Y	-Y
Total		-X	X	N

“n” representa el **total** de unidades de análisis incluidas en la muestra, lo que se suele denominar “la frecuencia de **orden cero**”. Por su parte, “-X”, “X”, “Y” y “-Y” son las frecuencias marginales o de **primer orden**; así, por ejemplo, “-X” representa el número total de casos que no presenta el atributo X, independientemente de que posean o no el atributo Y. Por último, “-XY”, “XY”, “-X-Y” y “X-Y” representan las **frecuencias condicionales**, o de **segundo orden**; de este modo, “-XY” significa el número absoluto de observaciones que combinan la ausencia del atributo X con la presencia de Y. Es importante notar que:

$$\begin{aligned}
 n &= (X) + (-X) \\
 &= (Y) + (-Y) \\
 &= (-X \ Y) + (X \ Y) + (-X-Y) + (X-Y)
 \end{aligned}$$

Aplicación del modelo a un ejemplo

Este modelo puede aplicarse para cualquier población y todo tipo de variables.

Tabla 4.1.: Misiones, 1980 – Población según tipo de asentamiento y pertenencia a hogares con Necesidades Básicas Insatisfechas

Tipo de asentamiento	Hogares con NBI		Total
	No	Sí	
Urbano	194.397	96.610	291.007
Rural	122.701	166.814	289.515
Total	317.098	263.424	580.522

Fuente: elaboración propia, en base a datos de Argentina, 1984: 343.

Así, en esta tabla, las unidades de análisis son personas, y los atributos “X” e “Y” se traducen, respectivamente, en el hecho de pertenecer o no a un hogar con Necesidades Básicas Insatisfechas (NBI), y de residir en una zona urbana o rural (es decir, “no-urbana”). El título ya nos informa que se trata de la **población** de Misiones en 1980; el n corresponde por tanto a 580.522 **personas**. Se verifica efectivamente que:

$$580.522 = 317.098 + 263.424$$

$$= 291.007 + 289.515$$

$$= 194.397 + 96.610 + 122.701 + 166.814$$

También se cumple que: $291.007 = 194.397 + 96.610$, es decir que cada marginal es igual a la suma de las frecuencias condicionales en la hilera –o la columna– correspondientes. **El primer paso de cualquier análisis es verificar si la tabla “cierra”**, vale decir, si se cumplen las relaciones aritméticas que debe satisfacer cada cifra; en caso contrario es evidente que se ha producido algún error en la tabulación.

¿Qué puede afirmarse en base a esta Tabla 4.1.? Puede decirse que en Misiones, en el año 1980, había 580.522 personas; proposición que, a pesar de su veracidad, tiene el inconveniente de no hacer uso de toda la información contenida en dicha tabla.

Un modo de comenzar el análisis es describiendo los marginales. Así, se puede decir que de estos 580.522 habitantes de Misiones, 291.007 residían en zonas urbanas en tanto que 289.515 lo hacían en áreas rurales. Esto ya es más interesante, aunque la misma conclusión podría haberse derivado de una distribución de frecuencias simple:

Tabla 4.1. a : Misiones, 1980 – Población según tipo de asentamiento

Tipo de asentamiento	Número de habitantes
Urbano	291.007
Rural	289.515
Total	580.522

Fuente: Tabla 4.1.

¿Para qué sirven los porcentajes?

Vemos que hay más habitantes urbanos que rurales; exactamente, los primeros superan a los segundos en 1.492 personas. ¿Ahora bien, es ésta una diferencia importante? Depende del contexto dentro del cual ubiquemos ese número. Lo sería sin duda si comparáramos la cifra con los datos del Censo de Población de 1970 en el que los rurales aventajaban a los urbanos en 132.886 habitantes. Pero, considerando intrínsecamente los datos de la Tabla 4.1.a., la manera de apreciar la importancia de esas 1.492 personas de diferencia es poniéndolas en relación con el total de la población provincial.

Tabla 4.1.a.1.: Misiones, 1980 – Población según tipo de asentamiento (%)

Tipo de asentamiento	Número de habitantes	%
Urbano	291.007	50,1
Rural	289.515	49,9
Total	580.522	100,0

Fuente: Tabla 4.1.a.

Es decir, considerar el peso relativo de cada grupo sobre el total de población. Se observa así que la diferencia entre la población urbana y la rural es muy escasa: 50,1 a 49,9%. Hemos calculado el porcentaje de población urbana mediante la siguiente operación:

$$\frac{291.007 \times 100}{580.522} \quad (\text{o, en general : } \frac{Y \times 100}{n})$$

¿Cuándo redondear los porcentajes?

En realidad, el resultado de la operación aritmética anterior arroja la cifra de 50,12850503, la que nosotros hemos redondeado a un decimal anotando 50,1.³ ¿Por qué este redondeo? El interés de los porcentajes es indicar con la mayor claridad las dimensiones relativas de dos o más números, transformando a uno de esos números, la **base**, en la cifra 100. Es indudable que:

$$\frac{291.007}{580.522} = \frac{50}{100} = 50\%$$

Matemáticamente, éstas son expresiones equivalentes –o casi- pero es evidente que en un sentido psicológico “50%” es la manera más concisa, sencilla y ventajosa de denotar la relación que nos interesa. Si se conservan muchos decimales, sólo se logra tornar más engorrosa la lectura de la tabla y se pierde la ventaja de expresar las cifras en porcentajes. Por ende se puede recomendar siempre que sea ello posible, **como regla general, prescindir totalmente de los decimales**. “50,12850503” parece más preciso que “50%”; más es ésta una precisión engañosa,⁴ y que a todos los efectos prácticos o teóricos carece absolutamente de significado.⁵

Sin embargo, en la Tabla 4.1.a.1. hemos consignado “50,1%” y no “50%”. ¿Por qué ya esta primera infracción a la regla que acabamos de formular? En el caso que nos ocupa, existirían al menos dos posibles justificaciones: a) trabajando con una población de 580.522 personas, cada punto del primer decimal representa 580 individuos, una cantidad relevante para muchos propósitos; y b) porque si no conserváramos el primer decimal, obtendríamos el mismo porcentaje para ambos sectores de la población, y tal vez no nos interese producir este efecto.⁶

El otro marginal de la Tabla 4.1. podría dar lugar a un análisis en un todo análogo. Se concluiría así que un 45,4% -o un 45%- de la población total de Misiones vivía en 1980 en hogares con NBI.

¿Cómo se lee una tabla de contingencia?

Siempre que se considera una tabla de contingencia es recomendable comenzar el análisis por las distribuciones univariadas de los marginales, para luego pasar al examen de las frecuencias condicionales, que nos permitirá aprehender el sentido peculiar de cada cruce de variables.

En un paso ulterior podríamos entonces hacer una lectura de cada una de las cifras contenidas en las celdas de la Tabla 4.1.:

1. 194.397 personas vivían en hogares sin NBI en áreas urbanas;
2. 96.610 lo hacían en hogares con NBI en áreas urbanas;
3. 122.701 pertenecían a hogares sin NBI en áreas rurales;
4. 166.814 habitaban en áreas rurales en hogares con NBI.

Todas estas proposiciones son **verdaderas**, en el sentido de que traducen con exactitud el significado de cada cifra; pero consideradas en conjunto constituyen una lectura puramente redundante de la información contenida en la Tabla 4.1., y no agregan

nada a lo que ésta ya está mostrando por sí misma. En general, cuando se analizan tabulaciones bi-variadas, el interés debe focalizarse en determinar si existe alguna **relación** entre las dos variables. En otros términos, partimos siempre de una **hipótesis**, más o menos explícita, acerca de la existencia o no de una relación entre las dos variables.

Modos alternativos de análisis

Hay básicamente dos modos de abordar el análisis de una tabla. De acuerdo con una distinción establecida por Zelditch (1959), se la puede analizar de manera asimétrica o simétrica. En el modo asimétrico, el interés está puesto en observar el efecto de una de las variables sobre la otra. Por lo contrario, en el análisis simétrico no se presupone que una variable funja como “**causa**” de la otra. Abordaremos sucesivamente estas dos alternativas, teniendo siempre presente que una tabla no es intrínsecamente simétrica o asimétrica, sino que esta distinción se limita al modo en que se decide encarar el análisis en función de los objetivos del investigador.

¿Qué es analizar una tabla asimétricamente?

El caso asimétrico: se plantea siempre que se elige considerar que una variable – la variable **independiente**- incide sobre la distribución de la obra –la variable **dependiente**. Hay tablas en las que cualquiera de las dos variables puede fungir como “causa “ de la otra. Aunque también suele ocurrir que se “imponga”, por así decirlo, el análisis en una determinada dirección.

La “regla de causa y efecto” o “primera regla de Zeizel” se aplica, en palabras de su autor, “siempre que uno de los dos factores del cuadro dimensional pueda considerarse como causa de la distribución del otro factor. La regla es que **los porcentajes deben computarse en el sentido del factor causal**”.⁷

¿Puede esta regla aplicarse a nuestra Tabla 4.1.? Responder positivamente a esta pregunta supondrá considerar, por ejemplo, que el tipo de asentamiento de la población “determina” o “condiciona” una probabilidad diferencial de pertenecer a un hogar con NBI. Esta hipótesis es plausible, si se tiene en cuenta que, por lo general, en nuestros países subdesarrollados el nivel de vida de las poblaciones rurales es inferior al de los habitantes urbanos.

¿Cómo se lee el título de una tabla?

El título ya es merecedor de algunas observaciones, en tanto ejemplifica un cierto código cuyas reglas debemos conocer si queremos comprender acabadamente el significado de la Tabla 4.1.1.:

- ① Las dos variables se encuentran claramente identificadas; se trata, respectivamente, de “Pertenencia de la población a hogares con Necesidades Básicas Insatisfechas” (que en el encabezamiento de las columnas figura como “Hogares con NBI”, y cuyos valores son “Sí” y “No”) y de “Tipo de asentamiento” (con los valores “Urbano” y “Rural”).

Tabla 4.1.1. : Misiones, 1980 – Pertenencia de la población a hogares con NBI, según tipo de asentamiento (%)

Tipo de asentamiento	Hogares con NBI		Total (100%)
	No	Sí	
Urbano	66,8	33,2	291.007
Rural	42,4	57,6	289.515
Total	54,6	45,4	580.522

Fuente: Tabla 4.1.

② Entre los nombres de las dos variables se intercala la preposición “según”; no hubiera sido incorrecto utilizar alguna otra preposición, como “por” o “de acuerdo”, pero cabe atender al orden en que se introducen los nombres de las variables. Tomando el “tipo de asentamiento” como variable independiente, ésta es introducida a continuación de la preposición “según”; en efecto, la presentación de los datos en la Tabla 4.1.1. apunta a destacar esta idea: **según** sea su tipo de asentamiento tenderán las personas a diferir en cuanto al valor mantenido en la variable dependiente.

③ El título finaliza con la expresión “(%)”; ello nos indica que las cifras consignadas en las celdas son porcentajes, y no frecuencias absolutas.⁸

④ En el encabezamiento de la última columna aparece la expresión “**Total (100%)**”. Esto quisiera expresar: a) que en dicha columna las cifras no son porcentajes sino frecuencias absolutas; y b) que las cifras absolutas de la columna fueron tomadas como base para calcular los porcentajes de las celdas.⁹

¿Cómo se lee una cifra porcentual?

Procedamos ahora a la lectura de la Tabla 4.1.1. Habiendo tomado “Tipo de asentamiento” como variable independiente, hemos en consecuencia calculado los porcentajes “en el sentido de esta variable, nuestro “factor causal”. Ello quiere decir que **las bases para el cálculo porcentual están dadas por el total de casos para cada valor de la variable independiente.**

En la celda superior izquierda de la tabla observamos “66,8”, y sabemos –por título- que la cifra corresponde a un porcentaje. La lectura correcta de esta cifra tiene lugar en dos pasos, cada uno de los cuales supone responder a una pregunta.

① Lo primera que debemos inquirir es “¿66,8% **de qué?** (o ¿de quiénes)”. La única respuesta correcta es: “del 100% constituido por los 291.007 habitantes urbanos”; es decir, buscamos primero en la tabla dónde está el 100% -en la primera hilera-, y dirigimos luego nuestra vista hacia el encabezamiento de dicha hilera leyendo: “Urbano”. Cumplimentado este primer paso, estaremos en condiciones de preguntarnos con éxito...

② ...” ¿**Qué sucede** con este 66,8%?” y podremos responder: “viven en hogares sin NBI”. A esta segunda pregunta respondimos simplemente dirigiendo nuestra atención hacia el encabezamiento de la columna: “Sí”.

Así, el significado de la primera celda puede expresarse:

“De todos los habitantes urbanos de Misiones, hay un 66,8% que pertenece a hogares sin NBI”.

Igualmente correcto sería escribir:

“Un 66,8% de la población urbana vive en hogares sin NBI”.

Es obvio, que existe una posibilidad cierta de optar por diferentes redacciones, pero lo fundamental es que la expresión literaria respete el significado de la cifra. Se puede pensar en dos grandes tipos de problemas que se plantean en la lectura de los porcentajes.

❶ Hablaremos de problemas “**lógicos**” cuando se produce una **falsa** lectura de la cifra porcentual. Estos errores devienen de una confusión acerca de la **base** sobre la cual está calculado el porcentaje. En cualquier tabla de doble entrada, existen potencialmente tres bases sobre las cuales es posible calcular los porcentajes, a saber,

- El total de la hilera: “291.007”, en nuestro ejemplo;
- El total de la columna, “317.098”; y
- El “total total”, el “n”: “580.522”.

Se comete un **error lógico** cuando un porcentaje es leído sobre una base que no fue la utilizada para calcularlo. Así, si se lee “Un 66,8% de los habitantes de Misiones son urbanos y viven en hogares sin NBI”, la expresión lingüística da a entender que el porcentaje fue calculado sobre el total de la población provincial, con lo cual el enunciado pasa a expresar una proposición **falsa** (el porcentaje que correspondería a dicha expresión lingüística no sería “66,8” sino “33,5”). Igualmente erróneo sería escribir “En Misiones, un 66,8% de las personas pertenecientes a hogares sin NBI residen en asentamientos urbanos”. La construcción de esta frase supone que el 66,8% fue calculado sobre el total de personas pertenecientes a hogares sin NBI, con lo que el enunciado es también falso (para esta redacción, el porcentaje correcto sería “61,3”). Por ende, hay una sola manera de generar enunciados verdaderos; es eligiendo una construcción lingüística que dé cuenta sin ambigüedad alguna del modo en que la cifra porcentual ha sido efectivamente calculada.¹⁰

❷ Pero también se presentan problemas pragmáticos. Sucede que diferentes redacciones son susceptibles de comunicar distintos significados. Comparemos los siguientes enunciados:

- a) “**Más de** dos tercios de los habitantes urbanos viven en hogares que no presentan NBI”; y
- b) “**Solamente** un 66,8% de los habitantes urbanos pertenece a hogares sin NBI”.

Tanto “a” como “b” expresan correctamente el porcentaje, desde una perspectiva puramente lógica. Sin embargo, es evidente que ambos enunciados no tienen el mismo significado: ciertamente “a” trasunta una visión de la situación más optimista que “b”. Sucede que, como lo explicara hace tiempo el lingüista Roman Jakobson, no es posible denotar sin connotar: las operaciones de selección y combinación que están necesariamente en obra en la producción de todo discurso introducen en él una dimensión ideológica.¹¹ Podemos probar de eliminar los adverbios en nuestros enunciados “a” y “b”, con lo que obtenemos expresiones cuyo valor lingüístico es muy similar:

- a1) “Dos tercios de los habitantes urbanos viven en hogares que no presentan NBI”, y
- b1) “Un 66,8% de los habitantes urbanos pertenece a hogares sin NBI”.

Aparentemente habríamos eliminado así toda valoración, permaneciendo sólo la fría cifra. Pero esto es creer que el significado de un enunciado individual sólo depende de su contenido intrínseco. Lo cierto es que este enunciado se inserta en un contexto más amplio, el discurso al que pertenece, cuyo significado global concurre a producir, pero que a la vez determina grandemente su propia significación. Lo expresado abona la idea de que estos problemas que hacen a la pragmática del discurso son inevitables. A lo sumo puede intentarse limitarlos controlando en alguna medida la adverbicación y la adjetivación.

¿Cómo se lee un conjunto de porcentajes?

Podemos ahora leer el conjunto de las cifras de la Tabla 4.1.1., que traducimos en la siguiente serie de enunciados:

1. Un 66,8% de los habitantes urbanos pertenece a hogares sin NBI;
2. Un 33,2% de los habitantes urbanos pertenece a hogares con NBI;
3. Un 42,4% de los habitantes rurales pertenece a hogares sin NBI;
4. Un 57,6% de los habitantes rurales pertenece a hogares con NBI;
5. Un 54,6% de todos los habitantes pertenece a hogares sin NBI; y
6. Un 45,4% de todos los habitantes pertenece a hogares con NBI.

Todos estos enunciados son verdaderos. Empero, su mera enumeración no constituye una “buena” lectura de la Tabla 4.1.1. En efecto, este conjunto de enunciados a) es en gran medida redundante; y, sobre todo, b) no apunta a destacar lo fundamental, esto es, la relación entre las variables que postula nuestra hipótesis y que es la única razón por la que los datos han sido presentados como se lo ha hecho, calculando los porcentajes en una dirección determinada.

En cuanto a la redundancia, debe resultar claro que el contenido del enunciado ya está incluido –implícitamente– en el enunciado 1: si un 66,8% de los habitantes urbanos pertenece a hogares sin NBI, y nos encontramos tratando con una variable dicotómica, ello implica que **necesariamente** hay un 33,2% de los habitantes urbanos en hogares con

NBI.¹² Y viceversa: si es verdadero el enunciado 2, necesariamente lo será también el 1. Es evidente que la misma relación se da para los pares de enunciados 3 y 4, y 5 y 6.

Aunque ello no resulte tan obvio, también son redundantes en cierto modo los porcentajes correspondientes a la hilera del total. Así, el que 45,4% de todos los habitantes pertenezcan a hogares con NBI no es más que el resultado de un promedio ponderado entre el 33,2% de urbanos y el 57,6% de rurales que presentan esta característica. Es ésta una propiedad interesante de los porcentajes marginales: necesariamente su valor se ubicará dentro de un rango limitado por los valores porcentuales consignados en las celdas correspondientes; en este caso, el porcentual del marginal deberá ser superior a 33,2 e inferior a 57,6, el que se encuentre “más cerca” de una y otra de estas cifras dependerá sólo del peso relativo de ambos grupos (el “rural” y el “urbano”) sobre la población total.¹³ Es por esta razón que frecuentemente se omite la presentación de los porcentajes marginales.¹⁴

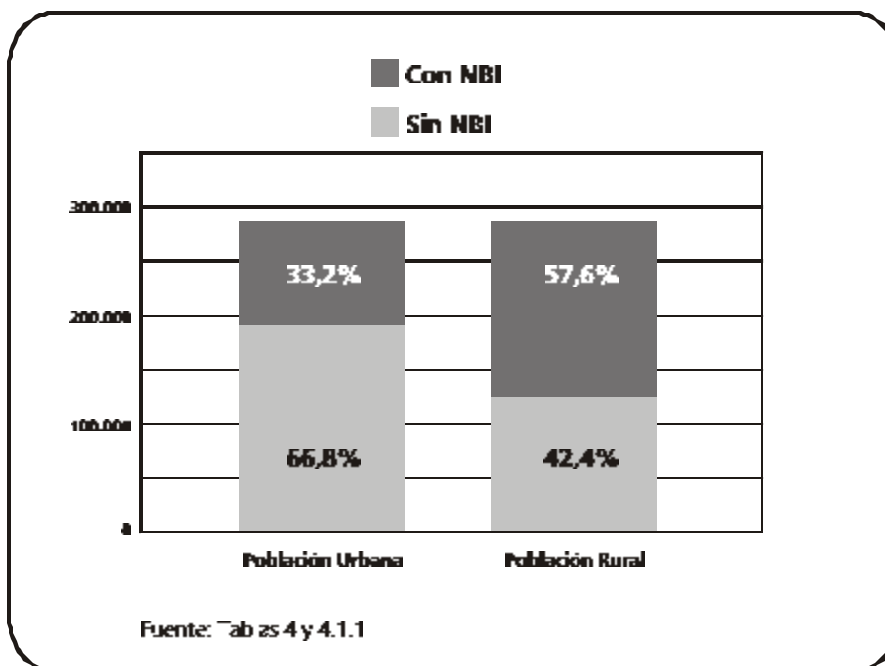
En suma, si intentamos reducir al mínimo la redundancia en la lectura de la tabla, podemos considerar que lo esencial de la información está contenido en los enunciados 2 y 4 (o, indiferentemente, en los 1 y 3). De este modo, destacaremos el sentido fundamental que queremos prestarle a los datos: en estas dos cifras –33,2% y 57,6%–¹⁵ está resumido lo que la tabla significa para nosotros. Comparando estos dos porcentajes, nuestra lectura pone en evidencia la relación estocástica entre las dos variables postulada por nuestra hipótesis:

“Mientras que en la población urbana hay un 33,2% de habitantes en hogares con NBI, entre los pobladores rurales este porcentaje asciende al 57,6%”.

Se corrobora por lo tanto la existencia de una probabilidad diferencial de pertenecer a un hogar con NBI en función del tipo de asentamiento de la población.

Una alternativa interesante para presentar esta información puede ser mediante un gráfico de columnas. Se ve claramente cómo ambas poblaciones son de tamaños similares y cómo la proporción de personas pertenecientes a hogares con NBI es mucho mayor en el campo.¹⁶

Figura XVI: Misiones, 1980 – Distribución de la población urbana y rural según su pertenencia a hogares con Necesidades Básicas Insatisfechas



Primera regla de Zeizel

Nos encontramos ahora en condiciones de completar nuestra formulación de la primera regla de Zeizel, referida al caso del análisis asimétrico:

“LOS PORCENTAJES SE CALCULAN EN EL SENTIDO DE LA VARIABLE INDEPENDIENTE, Y SE COMPARAN EN EL SENTIDO DE LA VARIABLE DEPENDIENTE.”¹⁷

En efecto, esto es todo lo que hemos hecho: hemos computado los porcentajes en el sentido horizontal (en este caso), y los hemos leído en el sentido vertical.¹⁸ Lo fundamental es calcular los porcentajes en la dirección adecuada, esto es, en el sentido de la variable a la que asignaremos el rol de “independiente” en nuestro análisis.

¿Cuál es la variable independiente?

Ahora bien, cuál sea la variable independiente, es una cuestión que está enteramente supeditada a los objetivos de nuestro análisis. Así como en la Tabla 4.1.1. se eligió el “Tipo de asentamiento”, así puede tomarse también como independiente la variable “Pertenencia a hogares con NBI”:

Tabla 4.1.2.: Misiones, 1980 – Tipo de asentamiento de la población según pertenencia a hogares con Necesidades Básicas Insatisfechas (%)

Tipo de asentamiento	Hogares con NBI		Total 100%
	No	Sí	
Urbano	61	37	50
Rural	39	63	50
Total	(317.098)	(263.424)	(580.522)

Fuente: Tabla 4.1.

¿En qué sentido puede pensarse que el hecho de pertenecer o no a un hogar con NBI “determine” el tipo de asentamiento de las personas? ¿Puede sostenerse una brumosa hipótesis según la cual los “pobres” preferirían residir en áreas rurales? La Tabla 4.1.2. nos permite pensar la diferencia que media entre las expresiones “variable independiente” y “causa”. Es evidente que en este caso no tiene demasiado sentido pensar en la pobreza como “causa” del tipo de asentamiento.¹⁹ Pero es perfectamente posible leer:

“En las zonas rurales de la Provincia se concentra el 63% de las personas pertenecientes a hogares con NBI, frente a sólo un 39% de las que pertenecen a hogares no carenciados”.

Se trata de una presentación de los datos de la Tabla 4.1. que tiende a destacar cómo la pobreza se concentra mayoritariamente en las áreas rurales de Misiones. No solamente la Tabla 4.1.2. es tan “verdadera” como la 4.1.1., sino que ambas son igualmente válidas. Aún cuando la Tabla 4.1.1. indujera en nosotros un mayor sentimiento de satisfacción, esta segunda interpretación no sería menos legítima por ello.

Segunda regla de Zeizel

Existe sin embargo una limitación al sentido en que es lícito computar los porcentajes, cuando se trabaja con datos muestrales. No siempre las muestras tienen la virtud de ser autoponderadas. Por diversas razones, puede ocurrir que una muestra no sea representativa de la población en algún sentido; hemos visto en el capítulo anterior que el caso es frecuente al utilizar diseños de muestra estratificados o por cuotas.

Imaginemos que queremos investigar acerca de la conformidad de un grupo de estudiantes de las carreras de Trabajo Social y de Turismo con el sistema de promoción por examen final; a tales efectos, seleccionamos una muestra de 40 alumnos de cada carrera, a sabiendas de que los totales de alumnos eran de 160 para Trabajo Social y de 80 para Turismo.

Tabla 4.2.: Conformidad con el sistema de examen final según carrera

Conformidad con el examen final	Carrera		Total
	Turismo	Trabajo Social	
Sí	7 18%	16 40%	23 29%
No	33 82%	24 60%	57 71%
Total	40 100%	40 100%	80 100%

Fuente: elaboración propia.

En casos semejantes sólo cabe calcular los porcentajes en el sentido en que se lo ha hecho en el ejemplo de la Tabla 4.2., y se podrá concluir que la proporción de disconformes con el sistema de aprobación por examen final es más elevada entre los alumnos de Turismo (82%) que entre los de Trabajo Social (60%).

En general, la segunda regla de Zeisel –que no es más que una limitación a la primera- afirma:

“CUANDO UN CONJUNTO DE MARGINALES NO ES REPRESENTATIVO DE LA POBLACIÓN, LOS PORCENTAJES DEBEN COMPUTARSE EN LA DIRECCIÓN EN QUE LA MUESTRA NO ES REPRESENTATIVA”.²⁰

En efecto, en nuestra muestra la relación entre los alumnos de las dos carreras es de 1:1 (40 en cada una), en tanto sabemos que en la población la relación real es de 1:2 (hay el doble de alumnos en Trabajo Social). Como nuestra muestra no es representativa por carrera, los porcentajes sólo pueden calcularse en esa dirección: sobre el total de alumnos de cada carrera.

¿Qué ocurriría si calculáramos directamente los porcentajes en el sentido horizontal? Concluiríamos –erróneamente- que del total de los estudiantes que se manifiestan conformes con el sistema de examen final hay un 70% que pertenece a la carrera de Trabajo Social:

Tabla 4.2.1. : Carrera según conformidad con el sistema de examen final (%)

Conformidad con el examen final	Carrera Turismo	Trabajo Social	Total
Sí	30	70	100
No	58	42	100

(n = 80)

Fuente: Tabla 4.2.

Es verdad que **en la muestra** se da este 70%; pero ello ocurre debido a un factor arbitrario que es el tamaño relativo de la muestra en ambas carreras. Como en la muestra la carrera de Trabajo Social se encuentra subrepresentada con relación a su peso real en la población, y sus estudiantes son más conformistas que los de Turismo, en la población deberá ser mayor el porcentaje de conformes concentrados en aquella carrera.

Supongamos que de haber trabajado con el universo, se hubieran obtenido las mismas proporciones de conformistas en ambas carreras que las registradas en la Tabla 4.2. Los resultados serían los presentados en la Tabla 4.2.2.²¹

Tabla 4.2.2.: Carrera según conformidaad con el examen final

Conformidad con el examen final	Carrera Turismo	Trabajo Social	Total
Sí	14 18%	74 82%	68 100%
No	66 41%	96 59%	162 100%
Total	80 33%	160 67%	240 100%

Fuente: Elaboración propia.

Se observa que hay en realidad un 82% de los conformistas que pertenecen a Trabajo Social y que, por lo tanto, el 70% que arrojaba la Tabla 4.2.1. no podía ser tomado como una estimación válida de la proporción existente en la población . Al no ser representativa la muestra en cuanto al peso relativo de ambas carreras, el cómputo directo de los porcentajes sólo se puede realizar como se lo hizo en la Tabla 4.2. Si se desea calcular los porcentajes en la otra dirección, no se lo puede hacer directamente, sino que es indispensable recurrir a algún sistema de ponderación de las frecuencias análogo al utilizado en la Tabla 4.2.2.

¿Y el modo simétrico?

Las dos reglas de Zeizel sintetizan lo esencial para el tratamiento asimétrico de tablas de contingencia. El análisis simétrico de estas tablas reviste comparativamente un interés menor. En este caso se computarán los porcentajes correspondientes a todas las frecuencias condicionales y marginales sobre la misma base del total de casos. En el análisis asimétrico, el cálculo de los porcentajes sobre columnas –o sobre hileras– permite lograr una estandarización de las frecuencias condicionales que quedan así liberadas de los efectos de las diferencias marginales. Esto nos permitía en la Tabla 4.2. comparar un 82% de disconformes en Turismo con un 60% en Trabajo Social, aún cuando en la población hubiera el doble de Trabajadores Sociales. En cambio, si los porcentajes se calculan todos sobre el “n” del cuadro, no se logra ninguna estandarización, ya que las diferencias marginales continúan pesando sobre las frecuencias condicionales. En este sentido, debe resultar evidente la necesidad de que la muestra sea representativa en todos los sentidos, si se desea analizar simétricamente una tabla compuesta a partir de observaciones muestrales. Así, no cabría someter la Tabla 4.2 a un tratamiento simétrico, por la misma razón que tampoco resultaba lícito el cómputo horizontal de los porcentajes.

Pero, sobre todo, el análisis simétrico no es apto para examinar la existencia de una relación de dependencia entre las dos variables; optamos por este tipo de análisis cuando **no** interesa indagar acerca del presunto “efecto” de una variable sobre la otra. Así la Tabla 4.1. podría también ser analizada simétricamente:

Tabla 4.1.3. : Misiones, 1980 – Distribución de la población por tipo de asentamiento y pertenencia a hogares con NBI

Tipo de asentamiento	Hogares con NBI		Total
	NO	SÍ	
Urbano	33,5	16,6	50,1
Rural	21,1	28,8	49,9
Total	54,6	45,4	(580.522)

Fuente: Tabla 4.1.

Leeremos así que, de los 580.522 habitantes de la Provincia, hay un 33,5% que pertenece a hogares urbanos sin NBI, seguido por un 28,8% de rurales con NBI, 21,1% rurales sin NBI y 16,6% de urbanos con NBI. En esta forma de presentación de los datos, ya no se visualiza con la misma claridad el efecto de una variable sobre la otra, lo que no implica que ésta deje de ser una interpretación tan legítima como las anteriores. Simplemente, habrá variado nuestro propósito. Es posible, por ejemplo, que tengamos un interés especial en saber que un 28,8% de la población de Misiones pertenece a hogares rurales con NBI, para comparar esa cifra con el 1,8% que se registra para la misma categoría de población en la Provincia de Buenos Aires, más urbanizada y menos pobre, o con el 30,0% de la vecina Corrientes, más urbanizada y más pobre.

La Tabla 4.1. se basa en datos censales. Muchas investigaciones realizadas por muestreo pueden no perseguir el objetivo de determinar la existencia de una relación

entre dos variables, sino proponerse la simple estimación de la proporción de una población que reúne determinadas características. De ser el caso el tratamiento simétrico de los datos obtenidos por muestra permite obtener estimaciones de las proporciones de personas dentro de cada categoría de la población. Pero, si por el contrario el objetivo es establecer una relación de dependencia entre dos variables, convendrá tratar la tabla asimétricamente.

2. EL ANÁLISIS DE LA RELACIÓN ENTRE VARIABLES

Cuando observamos mediante el tratamiento asimétrico de una tabla que una de las variables aparece determinando o afectando a la otra, podemos decir que ambas variables están **asociadas**.²² La medida de asociación más frecuentemente utilizada es, por lejos, la diferencia porcentual. Por otra parte, cuando se trata con muestras se plantea el problema adicional de determinar la significación estadística que se le puede prestar a una asociación entre variables. Abordaremos sucesivamente estos aspectos, para presentar luego algunos coeficientes de asociación.

2.1. La diferencia porcentual: una medida de la asociación.

Por su simplicidad de cálculo y por la claridad de su significado, la diferencia porcentual es sin duda la medida de asociación más popular. En esencia consiste en una sistematización de la primera regla de Zeizel.

Consideremos el siguiente ejemplo, cuyos datos provienen de una muestra de 121 estudiantes,²³ partiendo de la hipótesis de que el grado de conocimiento político condiciona el grado de participación política.²⁴

Tabla 4.3.: Grado de participación política y grado de conocimiento político

Participación política	Conocimiento político		Total
	Bajo	Alto	
Alta	6	13	19
Baja	59	43	102
Total	65	56	121

Si lo que se quiere es comprobar el efecto del conocimiento sobre la participación, los porcentajes se deben computar en el sentido de la variable “conocimiento”, o sea verticalmente:

Tabla 4.3.1.: Participación política según conocimiento político (%)

Participación política	Conocimiento político		Dif. %
	Bajo	Alto	
Alta	9	23	+14
Baja	91	77	-14
Total	100	100	(n=121)

Hemos simplemente aplicado la regla según la cual, los porcentajes se computan en dirección de la variable independiente y se comparan en la otra dirección. Salvo que ahora hacemos aparecer en la última columna la diferencia porcentual:

LA DIFERENCIA PORCENTUAL SE CALCULA EN LA DIRECCIÓN EN QUE SE REALIZA LA COMPARACIÓN.

Mientras que en los alumnos de Bajo conocimiento sólo hay un 9% con alta participación, entre los de Alto conocimiento hay un 23%: es decir, hay un **14% más** de alta participación política.²⁵

Ahora bien, imaginemos que quisiéramos en cambio determinar el efecto de la participación sobre el conocimiento:

Tabla 4.3.2.: Conocimiento político según participación política (%)

Participación política	Conocimiento político		Total
	Bajo	Alto	
Alta	68	32	100
Baja	42	58	100
Dif. %	+26	-26	(n =121)

Entre los altamente participativos hay un **26% más** con alto conocimiento. Este 26% es **también** una medida de la asociación entre las variables, y tan válida como la anterior, aunque a todas luces diferente. Según sea nuestro interés, podremos optar por una y otra cifra; pero lo que muestra el ejemplo es que la diferencia porcentual no nos brinda una medida **general** de la asociación en la tabla. Sucede que los porcentajes son sensibles a los cambios en las distribuciones marginales, y que precisamente en la Tabla

4.3. estos marginales difieren en forma notable (19 y 102 para “participación”; 56 y 65 para “conocimiento”).

Cualquier uso de la diferencia porcentual como indicador resumen de la asociación en una tabla implica una gran pérdida de información. Además, este problema se magnifica al trabajar con tablas de formato mayor al 2 x 2; cuanto más elevado sea el número de valores de cada variable, se multiplicará la cantidad de diferencias porcentuales computables, y resultará aún más discutible la elección de una de las tantas diferencias posibles como medida resumen de la asociación en la tabla.

La Tabla 4.4. permite ilustrar este problema.²⁶ Puesto que el NES ha sido tomado como variable independiente, lo lógico es leer el cuadro comparando entre sí los porcentajes de cada hilera. Evidentemente, el sentido general de la tabla es que cuanto menor es el NES, más negativa resulta la evaluación de la situación: ello surge nítidamente de la comparación de los porcentajes de la última hilera. Pero es claro que no existe una única diferencia porcentual, sino nueve posibilidades distintas de cómputo de diferencias; solamente en esa última hilera, sería posible comparar 45 con 30, 30 con 24, o 45 con 24; y va de suyo que ninguna de estas diferencias es más “verdadera” que las otras.

Tabla 4.4.: Evaluación de la situación social según NES (%)

Evaluación de la situación	Nivel Económico-Social			Total
	Bajo	Medio	Alto	
Favorable	37	38	47	41
Neutra	18	32	29	27
Desfavorable	45	30	24	32
Total (100%)	(40)	(73)	(49)	(162)

Fuente: Encuesta sobre comportamiento electoral, 1987.

Aún tomando en consideración estos defectos, no cabe menospreciar a la diferencia porcentual como instrumento del análisis. De hecho, en su práctica cotidiana el investigador la aplicará casi instintivamente, para tener una medida rápida de la asociación. Por lo demás, al trabajar con muestras la cuestión no radica simplemente en determinar el grado en que dos variables están asociadas, sino que se plantea un problema adicional.

¿Es esta relación estadísticamente significativa?

En la Tabla 4.3.1. la hipótesis inicial parecía corroborarse. Se observaba en efecto una diferencia positiva del 14% en cuanto a la participación de los estudiantes que contaban con un mayor grado de conocimiento político. Sin embargo, esta relación se

verifica en una muestra, constituida por 121 estudiantes que eran sólo una parte de la totalidad de los estudiantes de la FHCS-UnaM en 1984. La muestra con la que trabajamos es solamente una de las tantas muestras que se hubieran podido extraer del universo de la investigación. Tal vez el azar haya sido la razón de que apareciera en la muestra este 14% más, cuando en realidad esta relación no se daba en el universo. La cuestión es: ¿Podemos considerar a esa diferencia del 14% lo suficientemente importante como para asumir que representa una diferencia existente realmente en el universo? Cuando nos formulamos este tipo de preguntas, estamos inquiriendo si la relación es **estadísticamente significativa**.

¹ Por cierto, la tabla de contingencia puede también utilizarse para volcar datos provenientes de mediciones realizadas en el nivel intercalar, pero a costa de una gran pérdida de información; como veremos, existen otras técnicas mucho más precisas, matemáticamente hablando, para el análisis de tales variables.

² En la era actual de difusión masiva de las computadoras personales, es francamente desaconsejable recurrir a modos manuales de tabulación o bien a antigüedades tales como las tarjetas tipo McBee, las máquinas clasificadoras de tarjetas tipo Hollerit, etc.

³ El neologismo “redondear” significa suprimir los decimales –números a la derecha de la coma-, o conservar una limitada cantidad de éstos. El redondeo se realiza observando el decimal siguiente al que se quiere conservar; en el ejemplo, el segundo decimal es un 2 –cifra comprendida entre 0 y 4- por lo que corresponde anotar “50,1”, mientras que, de tratarse de un número igual o superior a 5, se anotaría “50,2”.

⁴ Dada la imprecisión de nuestros instrumentos de medición.

⁵ Según Galtung, “La presentación de los porcentajes con 1 o incluso con 2 decimales no tiene sentido a menos que 1) la **calidad** de la recolección sea tan buena que tenga sentido decir que el 70,1% y no el 70,2% dicen “sí”, etc.; 2) el **propósito** de la recolección de datos sea tal, que sea diferente para la interpretación que el 70,1% y no el 70,2% diga “sí”, etc. En general sugerimos que los porcentajes deben presentarse sin ningún decimal, para evitar una impresión de exactitud que es a menudo completamente espuria” (1966: II. 231). G. Bachelard decía que “El exceso de precisión, en el reino de la cantidad, se corresponde muy exactamente con el exceso de pintoresco, en el reino de la cualidad”, y veía en tales excesos las marcas de un espíritu no científico (1972:212).

⁶ En una visión **diacrónica**, “50,1%” podría tener el valor de significar la inversión de una tendencia: de un predominio histórico de la población rural, se pasa a una preponderancia de los habitantes urbanos. En cambio, desde un punto de vista **sincrónico**, podría convenir escribir “50%”, destacando la semejanza cuantitativa entre ambos sectores.

⁷ Cf. Zeisel, 1962: 37.

⁸ Igualmente claro sería omitir el signo de porcentaje en el título y consignarlo a continuación de cada cifra: 66,8%, 33,2%, etc.

⁹ Esta última convención, además de ser tan arbitraria como las anteriores, está lejos de ser universalmente reconocida. Otra manera habitual de proceder es hacer figurar en la última columna para todas las hileras la expresión “100,0”, pero aparte de ser ésta una información redundante (en el primer renglón es obvio que $33,2 + 66,8 = 100,0$), este procedimiento tiene el inconveniente de que hace desaparecer toda referencia a las frecuencias absolutas que fueron tomadas como base; en cambio, mientras éstas continúan apareciendo siempre será posible reconstruir las frecuencias absolutas correspondientes a las celdas: por ejemplo, $291.000 \times 0,332 = 96.614 \approx 96.610$ (la pequeña diferencia deviene del redondeo del porcentaje). Otra posibilidad es anotar entre paréntesis las bases de los porcentajes: “00,0 (291.007)”. También es posible duplicar cada cifra del cuadro consignando siempre las frecuencias absolutas y relativas –estas últimas de preferencia con algún recurso tipográfico distinto-, aunque esta práctica tiende a restarle nitidez a los datos.

¹⁰ Puesto que siempre existen tres alternativas para el cálculo de los porcentajes, es un hecho tan lamentable cuan inevitable que para leer un porcentaje siempre existan dos posibilidades de equivocarse y sólo una de acertar...

¹¹ Verón caracterizó a la ideología “como un **nivel de significación** de todo discurso transmitido en situaciones sociales concretas, referido al hecho inevitable de que, por su propia naturaleza, todo mensaje transmitido en la comunicación social posee una dimensión connotativa” (Cf. Verón, 1972:309).

¹² 33,2 es el complemento necesario para alcanzar al 100,0%; en efecto, $100,0 - 66,8 = 33,2$.

¹³ En el caso particular de la tabla que nos ocupa, ambos grupos tienen aproximadamente el mismo peso, por lo que podría calcularse: $(33,2 + 57,6) / 2 = 45,4\%$.

¹⁴ Sin embargo, cuando se trabaja con tablas de mayores dimensiones –y no basadas en dicotomías–, el porcentaje marginal puede funcionar como un punto de referencia útil que facilite la atribución de un significado a los porcentajes en las celdas. Incluso en nuestra misma Tabla 4.1.1. podría decirse que, frente a un 45,4% del total de la población que pertenece a hogares con NBI, el 33,2% de los “urbanos” es **comparativamente bajo**.

¹⁵ O alternativamente, en el par: 66,8% y 42,4%.

¹⁶ Los gráficos tienen un gran poder de comunicación y son un recurso al que se puede apelar en la presentación de informes de investigación. Es probable que muchas personas percibirán mejor una relación expresada visualmente, aunque se pierda algo de precisión con respecto a los datos numéricos. En este caso, el agregado de los porcentajes dentro de cada columna permite conservar lo esencial de la información numérica.

¹⁷ Esta fórmula pertenece a Zelditch (1959). Galtung dice: “sacar siempre los porcentajes perpendicularmente a la dirección de la comparación” (1966: II, 233).

¹⁸ Algunos autores sostienen la conveniencia de presentar siempre la variable independiente en el encabezamiento del cuadro y la variable dependiente en las hileras, con lo que los porcentajes se computarían siempre en el sentido vertical. Hay un único fundamento razonable para esta práctica, y es el mantener la analogía con el tratamiento de variables cuantitativas graficadas en un diagrama de ejes cartesianos, en el que la x –variable independiente– aparece siempre en la abscisa.

¹⁹ Estrictamente, para poder hablar de una relación causal entre las variables X e Y se requiere contar con evidencia de tres tipos: a) variación concomitante de X e Y; b) precedencia temporal de X con respecto a Y; y c) eliminación de otros posibles determinantes de Y (Cf. Sellitz et al., 1968: 100 y ss.).

²⁰ La expresión de la regla pertenece a Zelditch (1959).

²¹ Para construir la Tabla 4.2.2., simplemente multiplicamos por 2 las frecuencias absolutas correspondientes a los estudiantes de Turismo, y por 4 las de Trabajo Social.

²² El concepto de “asociación” se usará para describir la existencia de una relación entre variables en una tabla de contingencia; para distribuciones multivariantes se hablará de “correlación”.

²³ Los datos provienen del estudio “Participación política del estudiante de la FHCS – UnaM” (Inédito) realizado por alumnos de Antropología Social en 1984.

²⁴ Como habrá de verse a la brevedad, es igualmente plausible sostener la hipótesis de que “A mayor grado de participación política, mayor conocimiento”. En términos de Zetterberg, éste es un ejemplo de relación **reversible e interdependiente** entre las variables (1968: 59 y ss.).

²⁵ Igualmente podríamos haber comparado los porcentajes de Baja participación, encontrando que en los alumnos de Alto conocimiento hay un 14% menos. Tratando con variables dicotómicas, y considerando que por definición los porcentajes deben sumar 100%, no puede sorprendernos que las diferencias porcentuales sean en ambos renglones idénticas aunque de signo contrario: necesariamente, todo aumento del porcentaje en una categoría debe implicar una disminución en la otra.

²⁶ Los datos están tomados de un estudio (inédito) realizado por alumnos de la carrera de Antropología Social, en ocasión de las elecciones del 6 de setiembre de 1987 en Posadas.